Microbial Gene Essentiality

METHODS IN MOLECULAR BIOLOGY™

John M. Walker, SERIES EDITOR

436. Avian Influenza Virus, edited by *Erica Spackman, 2008*
435. Chromosomal Mutagenesis, edited by *Greg Davis and Kevin J. Kayser, 2008*
434. Gene Therapy Protocols: *Volume 2: Design and Characterization of Gene Transfer Vectors,* edited by *Joseph M. LeDoux, 2008*
433. Gene Therapy Protocols: *Volume 1: Production and In Vivo Applications of Gene Transfer Vectors,* edited by *Joseph M. LeDoux, 2008*
432. Organelle Proteomics, edited by *Delphine Pflieger and Jean Rossier, 2008*
431. Bacterial Pathogenesis: *Methods and Protocols,* edited by *Frank DeLeo and Michael Otto, 2008*
430. Hematopoietic Stem Cell Protocols, edited by *Kevin D. Bunting, 2008*
429. Molecular Beacons: *Signalling Nucleic Acid Probes, Methods and Protocols,* edited by *Andreas Marx and Oliver Seitz, 2008*
428. Clinical Proteomics: *Methods and Protocols,* edited by *Antonio Vlahou, 2008*
427. Plant Embryogenesis, edited by *Maria Fernanda Suarez and Peter Bozhkov, 2008*
426. Structural Proteomics: *High-Throughput Methods,* edited by *Bostjan Kobe, Mitchell Guss, and Huber Thomas, 2008*
425. 2D PAGE: *Volume 2: Applications and Protocols,* edited by *Anton Posch, 2008*
424. 2D PAGE: *Volume 1: Sample Preparation and Pre-Fractionation,* edited by *Anton Posch, 2008*
423. Electroporation Protocols, edited by *Shulin Li, 2008*
422. Phylogenomics, edited by *William J. Murphy, 2008*
421. Affinity Chromatography: *Methods and Protocols, Second Edition,* edited by *Michael Zachariou, 2008*
420. Drosophila: *Methods and Protocols,* edited by *Christian Dahmann, 2008*
419. Post-Transcriptional Gene Regulation, edited by *Jeffrey Wilusz, 2008*
418. Avidin–Biotin Interactions: *Methods and Applications,* edited by *Robert J. McMahon, 2008*
417. Tissue Engineering, Second Edition, edited by *Hannsjörg Hauser and Martin Fussenegger, 2007*
416. Microbial Gene Essentiality: *Protocols and Bioinformatics,* edited by *Andrei L. Osterman and Svetlana Y. Gerdes, 2008*
415. Innate Immunity, edited by *Jonathan Ewbank and Eric Vivier, 2007*
414. Apoptosis in Cancer: *Methods and Protocols,* edited by *Gil Mor and Ayesha Alvero, 2008*
413. Protein Structure Prediction, Second Edition, edited by *Mohammed Zaki and Chris Bystroff, 2008*
412. Neutrophil Methods and Protocols, edited by *Mark T. Quinn, Frank R. DeLeo, and Gary M. Bokoch, 2007*
411. Reporter Genes for Mammalian Systems, edited by *Don Anson, 2007*
410. Environmental Genomics, edited by *Cristofre C. Martin, 2007*
409. Immunoinformatics: *Predicting Immunogenicity In Silico,* edited by *Darren R. Flower, 2007*
408. Gene Function Analysis, edited by *Michael Ochs, 2007*
407. Stem Cell Assays, edited by *Vemuri C. Mohan, 2007*
406. Plant Bioinformatics: *Methods and Protocols,* edited by *David Edwards, 2007*
405. Telomerase Inhibition: *Strategies and Protocols,* edited by *Lucy Andrews and Trygve O. Tollefsbol, 2007*
404. Topics in Biostatistics, edited by *Walter T. Ambrosius, 2007*
403. Patch-Clamp Methods and Protocols, edited by *Peter Molnar and James J. Hickman, 2007*
402. PCR Primer Design, edited by *Anton Yuryev, 2007*
401. Neuroinformatics, edited by *Chiquito J. Crasto, 2007*
400. Methods in Membrane Lipids, edited by *Alex Dopico, 2007*
399. Neuroprotection Methods and Protocols, edited by *Tiziana Borsello, 2007*
398. Lipid Rafts, edited by *Thomas J. McIntosh, 2007*
397. Hedgehog Signaling Protocols, edited by *Jamila I. Horabin, 2007*
396. Comparative Genomics, *Volume 2,* edited by *Nicholas H. Bergman, 2007*
395. Comparative Genomics, *Volume 1,* edited by *Nicholas H. Bergman, 2007*
394. Salmonella: Methods and Protocols, edited by *Heide Schatten and Abraham Eisenstark, 2007*
393. Plant Secondary Metabolites, edited by *Harinder P. S. Makkar, P. Siddhuraju, and Klaus Becker, 2007*
392. Molecular Motors: *Methods and Protocols,* edited by *Ann O. Sperry, 2007*
391. MRSA Protocols, edited by *Yinduo Ji, 2007*
390. Protein Targeting Protocols, Second Edition, edited by *Mark van der Giezen, 2007*
389. Pichia Protocols, Second Edition, edited by *James M. Cregg, 2007*
388. Baculovirus and Insect Cell Expression Protocols, Second Edition, edited by *David W. Murhammer, 2007*
387. Serial Analysis of Gene Expression (SAGE): *Digital Gene Expression Profiling,* edited by *Kare Lehmann Nielsen, 2007*
386. Peptide Characterization and Application Protocols, edited by *Gregg B. Fields, 2007*
385. Microchip-Based Assay Systems: *Methods and Applications,* edited by *Pierre N. Floriano, 2007*
384. Capillary Electrophoresis: *Methods and Protocols,* edited by *Philippe Schmitt-Kopplin, 2007*
383. Cancer Genomics and Proteomics: *Methods and Protocols,* edited by *Paul B. Fisher, 2007*
382. Microarrays, Second Edition: *Volume 2, Applications and Data Analysis,* edited by *Jang B. Rampal, 2007*
381. Microarrays, Second Edition: *Volume 1, Synthesis Methods,* edited by *Jang B. Rampal, 2007*
380. Immunological Tolerance: *Methods and Protocols,* edited by *Paul J. Fairchild, 2007*
379. Glycovirology Protocols, edited by *Richard J. Sugrue, 2007*
378. Monoclonal Antibodies: *Methods and Protocols,* edited by *Maher Albitar, 2007*
377. Microarray Data Analysis: *Methods and Applications,* edited by *Michael J. Korenberg, 2007*
376. Linkage Disequilibrium and Association Mapping: *Analysis and Application,* edited by *Andrew R. Collins, 2007*
375. In Vitro Transcription and Translation Protocols: Second Edition, edited by *Guido Grandi, 2007*
374. Quantum Dots: *Applications in Biology,* edited by *Marcel Bruchez and Charles Z. Hotz, 2007*
373. Pyrosequencing® Protocols, edited by *Sharon Marsh, 2007*
372. Mitochondria: *Practical Protocols,* edited by *Dario Leister and Johannes Herrmann, 2007*
371. Biological Aging: *Methods and Protocols,* edited by *Trygve O. Tollefsbol, 2007*
370. Adhesion Protein Protocols, Second Edition, edited by *Amanda S. Coutts, 2007*
369. Electron Microscopy: *Methods and Protocols, Second Edition,* edited by *John Kuo, 2007*
368. Cryopreservation and Freeze-Drying Protocols, Second Edition, edited by *John G. Day and Glyn Stacey, 2007*
367. Mass Spectrometry Data Analysis in Proteomics, edited by *Rune Matthiesen, 2007*
366. Cardiac Gene Expression: *Methods and Protocols,* edited by *Jun Zhang and Gregg Rokosh, 2007*
365. Protein Phosphatase Protocols: edited by *Greg Moorhead, 2007*
364. Macromolecular Crystallography Protocols: *Volume 2, Structure Determination,* edited by *Sylvie Doublié, 2007*
363. Macromolecular Crystallography Protocols: *Volume 1, Preparation and Crystallization of Macromolecules,* edited by *Sylvie Doublié, 2007*

METHODS IN MOLECULAR BIOLOGY™

Microbial Gene Essentiality: Protocols and Bioinformatics

Edited by

Andrei L. Osterman
Burnham Institute for Medical Research
La Jolla, California

Svetlana Y. Gerdes
Fellowship for Interpretation of Genomes (FIG)
Burr Ridge, Illinois

HUMANA PRESS TOTOWA, NEW JERSEY

999 Riverview Drive, Suite 208
Totowa, New Jersey 07512

www.humanapress.com

All papers, comments, opinions, conclusions, or recommendations are those of the author(s), and do not necessarily reflect the views of the publisher.

This publication is printed on acid-free paper. ∞
ANSI Z39.48-1984 (American Standards Institute) Permanence of Paper for Printed Library Materials.

Cover design by Sandy M.S. Wong

Cover illustration: Arrangement by Sandy M.S. Wong of images from genetic footprinting electrophoresis analysis of essential genes in Haemophilus influenzae as reported in Akerley et al., PNAS 99(2): 966-971.

For additional copies, pricing for bulk purchases, and/or information about other Humana titles, contact Humana at the above address or at any of the following numbers: Tel.: 973-256-1699; Fax: 973-256-8341; E-mail: orders@humanapr.com; or visit our Website: www.humanapress.com

10 9 8 7 6 5 4 3 2 1

e-ISBN: 978-1-59745-321-9

Library of Congress Control Number: 2007926773

Preface

We would like to use this opportunity to say a few words about how and why this book emerged. Our story goes back to the early 2000s at Integrated Genomics Inc., in Chicago, when we embarked on a project fostered by Michael Fonstein to establish a high-throughput approach to systematically probe the relative importance (contribution to fitness) of genes in *Escherichia coli* under a variety of growth conditions. Since Fred Blattner's and, later, Hirotada Mori's groups were pursuing the gene-by-gene knockout strategy, we chose to adopt a complementary "transposomics" approach. This technique, if successfully implemented in set conditions, could be expanded toward comparative studies in multiple conditions and potentially in other species of clinical and industrial importance. We were convinced that a *comparative approach* would become a key to the successful analysis of gene essentiality data, just as it had proved to be valuable in other genomic techniques. This triggered the idea for this book, which was to bring together various research groups that developed and applied a variety of techniques for genome-scale analysis of gene essentiality in diverse microorganisms. We believed that it would not only provide guidance for future studies but also further the establishment of comparative analysis of gene essentiality as an important addition to the Systems Biology toolbox.

This book sends a message to new investigators that gene essentiality technology already exists in various implementations, ready for immediate application to numerous fundamental and practical tasks. Despite remaining hurdles, many technical problems have already been addressed and resolved due to ingenuity and persistence of pioneering research groups, many of which have contributed to this book. Still, this technology is not yet available as an off-the-shelf service. Hence, this book provides researchers with a first-stop guide for choosing the most appropriate strategy for their planned essentiality studies. Experimental and computational aspects are equally important in genome-scale gene essentiality analysis, as in all other genomic technologies, and we attempted to reflect both of these aspects in the book.

We are deeply grateful to all contributors who agreed to share their valuable experience in developing and applying this revolutionary methodology. We hope that their efforts (as well as patience and tolerance during the entire time between inception and publication of this book) will be rewarded by the utility and impact of this volume on the anticipated rapid progress of gene essentiality studies. We would especially like to thank John Walker for his inspiration and guidance in preparation of this book and Cindy Cook for her valuable help with technical editing and handling the manuscripts.

Andrei L. Osterman
Svetlana Y. Gerdes

Contents

Introduction

Part I Experimental Protocols

IA Populational Genome-Wide Essentiality Screens

IB Systematic Collections of Knockout Mutants

IC Genome Minimization

ID Conditional Knockouts

Part II Bioinformatics

IIA Statistics

IIB Data Integration and Modeling

Contributors

BRIAN J. AKERLEY • *Department of Molecular Genetics and Microbiology, University of Massachusetts Medical School, Worcester, MA*

FREDERICK M. AUSUBEL • *Department of Molecular Biology, Massachusetts General Hospital, Boston, MA*

TOMOYA BABA • *Nara Institute of Science and Technology (NAIST), Graduate School of Biological Sciences, Ikoma, Nara, Japan*

TAEOK BAE • *The University of Chicago, Department of Microbiology, Chicago, IL,* and *Indiana University Northwest, Gary, IN*

GÁBOR BALÁZSI • *Department of Systems Biology-Unit 950, University of Texas M. D. Anderson Cancer Center, Houston, TX*

JEF D. BOEKE • *Department of Molecular Biology and Genetics, The High Throughput Biology Center, The Johns Hopkins University School of Medicine, Baltimore, MD*

OU CHEN • *McKusick-Nathans Institute of Genetic Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD*

ANGELA M. CHU • *Departments of Biochemistry, Stanford University School of Medicine, Stanford, CA*

BÁLINT CSÖRGŐ • *Institute of Biochemistry, Biological Research Center of the Hungarian Academy of Sciences, Szeged, Hungary*

KIRILL DATSENKO • *Department of Biological Sciences, Purdue University, West Lafayette, IN*

RONALD W. DAVIS • *Departments of Biochemistry and Genetics, Stanford University School of Medicine, Stanford, CA,* and *Stanford Genome Technology Center, Palo Alto, CA*

TAMÁS FEHÉR • *Institute of Biochemistry, Biological Research Center of the Hungarian Academy of Sciences, Szeged, Hungary*

JOCHEN FÖRSTER • *Fluxome Sciences A/S, Lyngby, Denmark*

ALLYN FORSYTH • *GHC Technologies, Inc., La Jolla, CA*

SVETLANA Y. GERDES • *Fellowship for Interpretation of Genomes (FIG), Burr Ridge, IL*

ELIZABETH M. GLASS • *Mathematics and Computer Sciences Division, Argonne National Laboratory, Argonne, IL*

ALEXANDER I. GRENOV • *Thermo Fisher Scientific, Madison, WI*

ZSUZSA GYŐRFY • *Institute of Biochemistry, Biological Research Center of the Hungarian Academy of Sciences, Szeged, Hungary*

MASAYUKI HASHIMOTO • *Division of Gene Research, Department of Life Science, Research Center for Human and Environmental Science, Shinshu University, Tokida, Ueda, Nagano, Japan*

CHRISTOPHER D. HERRING • *Mascoma Corporation, Lebanon, NH*

HSUAN-CHENG HUAN • *Institute of Bioinformatics, National Yang-Ming University, Taipei, Taiwan*

RAFAEL IRIZARRY • *Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD*

MICHAEL A. JACOBS • *Department of Medicine, University of Washington Genome Center, Seattle, WA*

YINDUO JI • *Department of Veterinary and Biomedical Sciences, University of Minnesota, St. Paul, MN*

ANDREW R. JOYCE • *Bioinformatics Program, University of California, San Diego, La Jolla, CA*

ILDIKÓ KARCAGI • *Institute of Biochemistry, Biological Research Center of the Hungarian Academy of Sciences, Szeged, Hungary*

JUN-ICHI KATO • *Department of Biological Sciences, Graduate School of Science and Engineering, Tokyo Metropolitan University, Minaminohsawa, Hachioji, Tokyo, Japan*

SUN CHANG KIM • *Department of Biological Sciences, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea*

KWAN SOO KO • *Asian-Pacific Research Foundation for Infectious Diseases (ARFID), Seoul, Korea,* and *Division of Infectious Diseases, Samsung Medical Center Sungkyunkwan University School of Medicine, Seoul, Korea*

IRENA KUKAVICA-IBRULJ • *Centre de Recherche sur la Fonction, Structure et Ingénierie des Protéines (CREFSIP), Pavillon Charles–Eugène Marchand, Biologie Médicale, Faculté de Médecine, Université Laval, Québec, Canada*

ANUJ KUMAR • *Department of Molecular, Cellular, and Developmental Biology, Life Sciences Institute, The University of Michigan, Ann Arbor, MI*

JAMES M. LANE • *Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA*

ROGER C. LEVESQUE • *Centre de Recherche sur la Fonction,Structure et Ingénierie des Protéines (CREFSIP), Pavillon Charles–Eugène Marchand, Biologie Médicale, Faculté de Médecine, Université Laval, Québec, Canada*

NICOLE T. LIBERATI • *Department of Molecular Biology, Massachusetts General Hospital, Boston, MA*

HIDEO MATSUDA • *Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Osaka, Japan*

PAMELA B. MELUH • *Department of Molecular Biology and Genetics, The High Throughput Biology Center, The Johns Hopkins University School of Medicine, Baltimore, MD*

TAKEYOSHI MIKI • *Department of Physiological Sciences and Molecular Biology, Fukuoka Dental College, Sawaraku, Fukuoka, Japan*

DOMINIQUE MISSIAKAS • *Department of Microbiology, The University of Chicago, Chicago, IL*

HIROTADA MORI • *Nara Institute of Science and Technology, Takayama, Ikoma, Nara, Japan,* and *Institute of Advanced Biosciences, Keio University, Tsuruoka, Yamagata, Japan*

JEFFREY P. MURRY • *Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA*

JENS NIELSEN • *Center for Microbial Biotechnology Biocentrum, Technical University of Denmark, Lyngby, Denmark*

HIRONORI NIKI • *Genetic Strains Research Center, National Institute of Genetics, Mishima, Shizuoka, Japan*

ANDREI L. OSTERMAN • *Burnham Institute for Medical Research, La Jolla, CA*

BERNHARD Ø. PALSSON • *Department of Bioengineering, University of California, San Diego, La Jolla, CA*

XUEWEN PAN • *Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX*

BRIAN D. PEYSER • *United States Army Medical Research Institute of Infectious Diseases, Fort Detrick, Frederick, MD*

GYÖRGY PÓSFAI • *Institute of Biochemistry, Biological Research Center of the Hungarian Academy of Sciences, Szeged, Hungary*

WILLIAM S. REZNIKOFF • *Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA*

ISABEL ROCHA • *Centro de Engenharia Biológica, Universidade do Minho, Campus de Gualtar, Braga, Portugal*

ERIC J. RUBIN • *Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA*

FRANÇOIS SANSCHAGRIN • *Centre de Recherche sur la Fonction, Structure et Ingénierie des Protéines (CREFSIP), Pavillon Charles–Eugène Marchand, Biologie Médicale, Faculté de Médecine, Université Laval, Québec, Canada*

CHRISTOPHER M. SASSETTI • *Department of Molecular Genetics and Microbiology, University of Massachusetts Medical School, Worcester, MA*

OLAF SCHNEEWIND • *Department of Microbiology, The University of Chicago, Chicago, IL*

MICHAEL D. SCHOLLE • *Amunix Inc., Mountain View, CA*

KAREN JOY SHAW • *Trius Therapeutics, Inc., San Diego, CA*

JAE-HOON SONG • *Asian-Pacific Research Foundation for Infectious Diseases (ARFID), Seoul, Korea,* and *Division of Infectious Diseases, Samsung Medical Center Sungkyunkwan University School of Medicine, Seoul, Korea*

SHARON SOOKHAI-MAHADEO • *Department of Molecular Biology and Genetics, The High Throughput Biology Center, The Johns Hopkins University School of Medicine, Baltimore, MD*

FORREST A. SPENCER • *Department of Molecular Biology and Genetics, McKusick-Nathans Institute of Genetic Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD*

TARA K. THURBER • *Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA*

CAROL TIFFANY • *McKusick-Nathans Institute of Genetic Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD*

KINGA UMENHOFFER • *Institute of Biochemistry, Biological Research Center of the Hungarian Academy of Sciences, Szeged, Hungary*

JONATHAN M. URBACH • *Department of Molecular Biology, Massachusetts General Hospital, Boston, MA*

LIANGSU WANG • *Merck & Co., Inc., Rahway, NJ*

XIAOLING WANG • *Department of Molecular Biology and Genetics, The High Throughput Biology Center, The Johns Hopkins University School of Medicine, Baltimore, MD*

BARRY L. WANNER • *Department of Biological Sciences, Purdue University, West Lafayette, IN*

OLIVER WILL • *Allan Wilson Centre, University of Canterbury, Christchurch, New Zealand*

KELLY M. WINTERBERG • *Department of Biology, University of Utah, Salt Lake City, UT*

SANDY M.S. WONG • *Department of Molecular Genetics and Microbiology, University of Massachusetts Medical School, Worcester, MA*

GANG WU • *Department of Molecular Biology, Massachusetts General Hospital, Boston, MA*

ZHIFANG XIE • *Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA*

YOSHIHIRO YAMAMOTO • *Department of Genetics, Hyogo College of Medicine, Nishinomiya, Hyogo, Japan*

YUKIKO YAMAZAKI • *Center for Genetic Resource Information, National Institute of Genetics, Mishima, Shizuoka, Japan*

DEZHONG YIN • *Aastrom Biosciences Inc., Ann Arbor, MI*

BYUNG JO YU • *Department of Biological Sciences, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea*

DANIEL S. YUAN • *Department of Molecular Biology and Genetics, The High Throughput Biology Center, The Johns Hopkins University School of Medicine, Baltimore, MD*

CHUN-TING ZHANG • *Department of Physics, Tianjin University, Tianjin, China*

REN ZHANG • *Department of Epidemiology and Biostatistics, Tianjin Cancer Institute and Hospital, Tianjin, China*

1

Overview of Whole-Genome Essentiality Analysis

Karen Joy Shaw

Genomic sequencing has transformed modern biology into an age of global analysis of gene expression, protein pathways, and metabolic networks. To understand whole-cell function, biologists and bioinformaticians in many fields have developed a diverse set of methods and tools to identify genes essential to a particular organism under a particular set of conditions. *Gene Essentiality: Protocols and Bioinformatics* reviews many of these diverse techniques and experimental procedures developed to analyze entire genomes of a variety of prokaryotes and eukaryotes.

The need to identify novel antibacterial and antifungal drug targets has been one of the major drivers for the development of techniques designed to determine gene essentiality. Through the large-scale identification of essential genes, an abundance of targets became available for drug screening. Many biotechnology and pharmaceutical companies spent a decade exploring essential gene research with the goal of identifying inhibitors of essential gene products that would mimic the phenotype of a gene knockout or knock-down. If an inhibitor could successfully reach the gene product, it would either kill or block the growth of any microbe that required the functional gene. Although there have been a few success stories, such as the peptide deformylase inhibitors, this approach has added few drugs to the antimicrobial arsenal. In general, the failure of this approach resulted from the inability to find inhibitors capable of permeating the cell, rather than from difficulty with inhibiting the particular target protein. The question, however, of which essential targets are “druggable” is still an open one and is somewhat negatively biased by the “anti-microbial-unfriendly” makeup of modern pharmaceutical small-molecule libraries.

One fundamental procedure for assessing gene essentiality is the use of transposon mutagenesis. Transposon insertion into a gene generally interrupts transcription; however, an insertion may also demonstrate polar effects on the transcription of distal genes in an operon. In addition, transposons sometimes have preferential sites of insertion. In **Chapter 2**, Reznikoff and Winterberg describe many of the transposon-based techniques that have been used to identify essential bacterial genes. In establishing these

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

techniques, several of these issues have been addressed, including the use of transposons containing an outward promoter to reduce polar effects on downstream gene expression. In addition, transposons have been developed or selected that randomly insert and are able to uniformly saturate the genome. Balázsi (**Chapter 23**) evaluates the validity of some of these assumptions, and Will (**Chapter 22**) discusses some of the statistical methods that are used in predicting the probability that a gene is essential from the data generated by random insertion libraries. Jacobs and Liberati et al. (**Chapters 9 and 10**) examine the development of random and near-saturation transposon insertions in *Pseudomonas aeruginosa*. In the former, a library of 30,100 unique transposon insertions was generated using Tn*5* IS*50*L; in the latter, an insertion library of 34,000 mutants was developed using a *mariner*-based transposon. Both of these libraries represent multiple insertions into nonessential genes. Bae et al. (**Chapter 7**) describe the development of a *mariner* transposon—based insertion library of *Staphylococcus aureus*. Miki et al. (**Chapter 13**) use mini-Tn*10* insertions to disrupt cloned fragments of the Kohara λ library of the *Escherichia coli* chromosome. λ lysogeny is used to transfer the disruptions into the *E. coli* chromosome by homologous recombination with the cloned insert, generating a partial diploid. A second recombination event generates either the wild type or the haploid disruption mutant, the latter of which is only recoverable for nonessential mutants.

Methods for analysis of transposon insertions can be categorized as either a gene-by-gene or a genomic approach. Genomic approaches generally involve assessing essentiality by "who is lost" from the population compared with a zero time (t_0) or compared with a population grown under another growth condition (e.g., minimal vs. rich media). Such techniques require enough generations of growth to dilute the signal of the mutagenized cell to the point where it is undetectable but fail to distinguish between overt cell death and the inability to grow under the particular selective condition (however viable). Although often functionally similar, distinguishing between these two results may have important ramifications in the search for drug targets that are likely to lead to bacteriostatic versus bactericidal agents.

Transposon mutagenesis has also been an important tool in the identification of essential genes in eukaryotes. Smith et al. *(1)* described an *in vivo* footprinting method in *Saccharomyces cerevisiae*, one of the first global strategies for simultaneous analysis of the importance of genes to the fitness of an organism under particular growth conditions. After insertional mutagenesis by a modified Ty1 element, the investigators divided the mutagenized population into aliquots that were each grown under different physiologic conditions. DNA was isolated at t_0 and after subsequent generations of growth, during which time there was a depletion of the mutated cells that were unable to grow (or grew more slowly). If a gene was essential, polymerase chain reaction (PCR) amplification of that gene using a gene-specific primer and a transposon-specific primer would result in fewer amplification products than the zero-time control. In collaboration with a team of scientists at Genome Therapeutics Corporation, my laboratory adapted this strategy to bacteria and developed a method for globally determining the importance of a particular gene to the fitness of *E. coli* using a mini-Tn*10* transposon with an outwardly oriented promoter. In addition, we demonstrated that genetic complementation of an essential gene restores the ability to detect PCR products from that gene *(2)*. Gerdes et al. *(3)* expanded on this technology and identified 620 essential genes and 3126 dispensable genes in *E. coli* under conditions of robust aerobic growth in rich media. Because there is a strong tendency of genes and functions that are defined as essential in *E. coli* to be essential throughout the bacterial kingdom, by presenting the full footprinting data set, this work has become a critical resource to scientists involved in antibacterial drug discovery as well as in basic research in bacterial physiology. This work is reviewed by Scholle and Gerdes (**Chapter 6**). Footprinting technology does have some limitations, including the difficulty in assessing the essentiality of small genes (<400 bp) due to the lower number of transposition events per gene, inability to assess duplicated genes or genes with functional paralogs, and regions that are "cold spots" for transposition.

Wong and Akerley (**Chapter 3**) review the techniques involved in the identification of essential genes using GAMBIT (genomic analysis and mapping by *in vitro* transposition) technology, which is similar to genetic footprinting but utilizes the *mariner* transposon to generate insertions into PCR-amplified genomic segments *in vitro*, with the subsequent introduction of these fragments into naturally competent bacteria (*Haemophilus influenzae* and *Streptococcus pneumoniae*). Although lacking a true t_0, essential genes can be identified and functional genomic analysis performed.

Similarly, Murry et al. (**Chapter 4**) describe a method using the *mariner* transposon in *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG that utilizes transposon site hybridization (TraSH). This technique can determine the complete set of genes required for growth under particular conditions by differential hybridization to microarrays but has the same limitations as the footprinting technology. TraSH has also been used to evaluate the requirements for *Mycobacterium tuberculosis* survival in a murine model of infection *(4)*.

Genes that are essential for pathogenesis formed another focus of inquiry in the study of infectious diseases and in the search for new antimicrobial agents. Techniques developed to identify genes critical for growth, survival, or virulence *in vivo* but not *in vitro* include *in vivo* expression technology (IVET), signature-tagged mutagenesis (STM), and microarray technology, all of which were adapted to a large variety of organisms. STM experiments involve transposon mutagenesis using an element that is engineered to contain a short variable region (the signature tag). After passing mutant pools through animals, the relative abundance of each tagged mutant in the input and output pools is compared, either by colony hybridization, dot blotting, hybridization to high-density oligonucleotide arrays, or by PCR. The latter adaptation of STM is described by Sanschagrin et al. (**Chapter 5**) and was used to identify *P. aeruginosa* genes that are critical in a rat model of chronic respiratory infection. One downside to this technique is that mutant collections must be grown in the laboratory prior to introduction into the animal, thus eliminating genes that are essential for growth/survival both *in vitro* and in an animal model. Additionally, gene targets identified through this methodology were, by definition, important for the establishment of infection (such as adherence, etc.) and would not necessarily be sensitive to small-molecule inhibitors once an infection was already established (e.g., maintenance).

Gene expression *in vivo* could also be assessed by IVET technology, which is used to detect genes that are transcriptionally induced during an infection. This technique identifies randomly cloned promoter sequences that permit expression of a promoterless gene required during bacterial growth in an animal host. Clones containing these "trapped promoters" are recovered and identified. One of the limitations of IVET is that the ability to detect a particular gene depends upon relative level and timing of gene expression *in vivo.* In addition, the number of organisms needed to be introduced to the host is relatively large, and therefore the gene regulation observed may not reliably reflect the events that occur during natural infection of the host.

In higher organisms, gene disruptions have been used for functional analysis, cellular network interpretation, and in the selection of targets for drug discovery. Kumar (**Chapter 8**) describes a methodology for functional analysis of *S. cerevisiae,* aided by the development of a single transposon designed for gene disruption, *lacZ* fusion, and epitope tagging. Yeast strains containing transposon insertions can be screened for phenotypes and/or protein localization, and the site of transposon insertion within these strains can be identified by PCR or other approaches. Most of the transposon insertion can be removed using cre-*lox* recombination, leaving behind an epitope tag. Chu and Davis (**Chapter 14**) report on the methodology used to create the publicly available yeast knockout collection containing deletions of nearly all of the approximately 6200 open reading frames in *S. cerevisiae*. Each deletion mutant is uniquely identified by a "molecular bar code" or tag, allowing parallel analysis of relative fitness under different physiologic growth conditions by microarray hybridization, reminiscent of footprinting and TraSH analysis. Meluh et al. (**Chapter 15**) describe the use of these *S. cerevisiae* knockouts in a diploid-based synthetic-lethality analysis by microarray (dSLAM) for the global identification of the fitness of double mutant strains and to monitor the genetic interactions between genes. Peyser et al. (**Chapter 25**) present statistical analysis methods for TAG microarray hybridization data to improve sensitivity and specificity.

Conditional lethal mutants and downregulated gene expression have often been used in high-throughput whole-cell screens to identify novel antimicrobial agents active against essential gene products. These screens operate under the premise that with diminished expression or activity, cells may be hypersensitive to compounds active against the gene product. Herring (**Chapter 21**) describes the identification and utility of conditional lethal amber mutations of *E. coli.* Similarly, antisense technology has also been used to identify essential genes, to evaluate their function during *in vitro* growth and during infections, and to screen compound libraries in cell-based comparative hypersensitivity assays. Antisense technology is based on the phenomenon that dsRNA is rapidly destroyed in organisms, leading to reduced gene expression. Yin and Ji (**Chapter 19**) and Forsyth and Wang (**Chapter 20**) describe the construction, characterization, and use of gene-specific and genomic antisense libraries in *S. aureus*.

In recent years, technologies for precise deletion of a gene have markedly improved in many organisms. Generally, they involve two recombination events: the first replaces the gene with a selectable marker, and the second removes the marker leaving minimal scarring at the site. The latter step is especially important in correctly assigning gene function in operons where there may be polar effects. Fehér et al. (**Chapter 16**) describe techniques that allow the scarless removal of single genes and the construction of serial deletions in *E. coli.* For nonessential genes, knockout mutants are an extremely useful tool for analyzing biological function and metabolic flux. Baba and Mori (**Chapter 11**) and Baba et al. (**Chapter 12**) describe the construction of the Keio collection of single gene deletions of all 3985 nonessential genes in *E. coli* K-12 and their utility for functional analysis. They report the inability to disrupt the remaining 303 genes, the hallmark of genes that are essential under the growth conditions tested. Allelic replacement mutagenesis is described by Song and Ko (**Chapter 28**) for *S. pneumoniae*.

Methods, such as precise deletions, have been critical for accurate assessment of the “minimal genome,” which has been estimated to be approximately 200 to 400 genes. Using Tn*5* transposons and a Cre/loxP excision system, Yu and Kim (**Chapter 17**) describe the construction of *E. coli* strains with large deletions. Through P1 transduction and recombination, a cumulative deletion strain was constructed that lacks nearly 300 open reading frames (ORFs) but exhibited normal growth. Kato and Hashimoto (**Chapter 18**) report a recombination method for preparing large-, medium-, and small-scale deletions and engineered an *E. coli* strain lacking approximately 30% of the genome.

The sequences of nearly 400 genomes are now publicly available, as are data analysis tools including ORF prediction, genomic comparisons, motif identification, and protein structural comparisons. Yamazaki et al. (**Chapter 26**) describe the PEC database of *E. coli* genes that includes data on essentiality, results of similarity searches, and information about structural domains, motifs, and homologues. Considerable progress has been made in gene annotation and the assignment of putative function; however, this continues to be an area of intensive work. A database of essential genes (DEG) is described by Zhang and Zhang (**Chapter 27**) that can be searched and BLASTed. DEG also includes functional information on the essential genes of nine genomes. It should be noted that discrepancies sometimes exist in the reported lists of essential genes for a particular organism ascertained by laboratories using different genome-scale techniques. Grenov and Gerdes (**Chapter 24**) discuss the basis for some of these differences.

Using bioinformatics to predict gene essentiality began with genomic comparisons to identify conserved gene families. After subtracting genes based on similarity to human or other mammalian databases, these families were often further parsed into “bacterial specific” or “fungal specific” gene lists, from which targets were chosen for antibacterial or antifungal drug discovery. However, assumptions about essentiality across species based on experimental evidence in one species are sometimes faulty due to gene duplications, gene substitutions, and alternative pathways. This occasional genetic diversity was found to significantly alter the anticipated bacterial spectrum of newly identified inhibitors of essential gene targets.

Better understanding of the metabolic capabilities of each particular organism is also critical to predicting conservation of essentiality across species. Rocha et al. (**Chapter 29**) describe a process of iterative modeling of metabolic networks that takes into account available literature, determinations of reaction stoichiometry, energy and metabolic flow, as well as other physiologic parameters. The resulting algorithms have important applications to engineering microbial metabolism for the production of desirable metabolites or for strain improvement. Joyce and Palsson (**Chapter 30**) use flux balance analysis, a mathematical technique, to assess the capabilities of metabolic networks using *E. coli* as a model. These *in silico* results are important in the interpretation of the complex relationship between genotype and phenotype and can be applied to our understanding and prediction of gene essentiality.

The sequence of a genome has become an important, basic tool for biologists interested in the function of a particular gene or set of genes and often provides insights in studies of metabolic pathways. In addition, comparative genomics and mutant analysis help to elucidate the role of specific genes in the life cycle and lifestyle of an organism. The technological advances in the evaluation of gene essentiality, either *in vitro* or *in vivo*, have resulted in the delineation of the critical genes for life in many different organisms. These findings contribute to our basic understanding of the biology of these organisms, our knowledge of host-pathogen relationships, and our strategies and directions for antimicrobial drug discovery, so critical in an era of increasing microbial resistance.

References

1. Smith, V., Botstein, D., and Brown, P. O. (1995) Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 6479–6483.
2. Hare, R. S., Walker, S. S., Dorman, T. E., Greene, J. R., Guzman, L. M., Kenney, T. J., et al. (2001) Genetic footprinting in bacteria. *J. Bacteriol.* **183**, 1694–1706.
3. Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balazsi, G., Ravasz, E., Daugherty, M. D., et al. (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684.
4. Sassetti, C. M, and Rubin, E. J. (2003) Genetic requirements for mycobacterial survival during infection. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12989–12994.

2

Transposon-Based Strategies for the Identification of Essential Bacterial Genes

William S. Reznikoff and Kelly M. Winterberg

Summary

We present a conceptual review of transposition-based strategies for determining gene essentiality on a one-by-one basis in bacteria. Many of the techniques are described in greater detail in individual chapters of this volume. The second section of this chapter deals with transposition-deletion—based strategies for determining the essentiality of blocks of genes. This latter approach has the potential to experimentally define the minimal required genome for a given organism.

Key Words: deletion; essential genes; insertion; transposon.

1. Introduction

A century of research work has been focused on the analysis of genetic determination of biological properties. With the advent of genome projects, in which the DNA sequences of the genomes for an ever-expanding group of organisms are now available, we are still faced with the daunting challenge of determining the functional importance of the various genes present in any given genome. One approach to this functional gene analysis is to determine which genes in an organism's genome are required for survival and growth in any particular environment; in other words, which genes are essential. A strategy for determining gene essentiality is to attempt an isolation of knockout mutations of the genes in question. Failure to isolate such a knockout mutation in a particular gene is taken as presumptive evidence that the gene in question is essential in the tested (all?) growth conditions. Alternatively, a gene might not be essential in one defined condition but be essential in another test circumstance. In these cases, the gene mutants can be studied for their effects on survival and growth under various test circumstances. DNA transposition, in which the transposon acts as an insertion mutagen or, in some cases, as a deletion mutagen, is a powerful approach for the generation of appropriate knockout mutations for these studies. This chapter provides an overview of transposition strategies for determining gene essentiality. The individual

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

strategies are described in more detail elsewhere in this volume and in other cited references.

There are two different types of questions that are addressed in these studies. The first most common approach is to ask whether a given particular gene is essential in an otherwise complete, intact genome. This one-by-one approach looks at particular genes but sometimes misses the particular functions encoded by the genes. This is because genomes sometimes contain more than one gene encoding products capable of performing the same function. We call such genes redundant. In this case, each such redundant gene could be individually destroyed with no impairment to the organism's survival and growth even if the function is essential. To determine that the particular function is essential, one would need to destroy all redundant genes and demonstrate survival and/or growth impairment. Thus, we must also look at strategies that can be used to define essential functions regardless of whether various functions are encoded by unique individual genes or redundant genes. For this type of inquiry, one can also use transposition-based approaches to generate large-scale deletions. These large-scale deletions not only offer an approach to identifying essential functions encoded by redundant genes but also suggest a strategy for dramatically shrinking the size of the organism's genome perhaps to the extent of defining a minimal essential genome.

This chapter shall first describe transposition systems that are used to generate individual insertion mutations. These techniques are based on the straightforward application of standard transposition mutagenesis that is schematically described in **Figure 1**. For the more global goal of shrinking the genome in order to define essential functions,

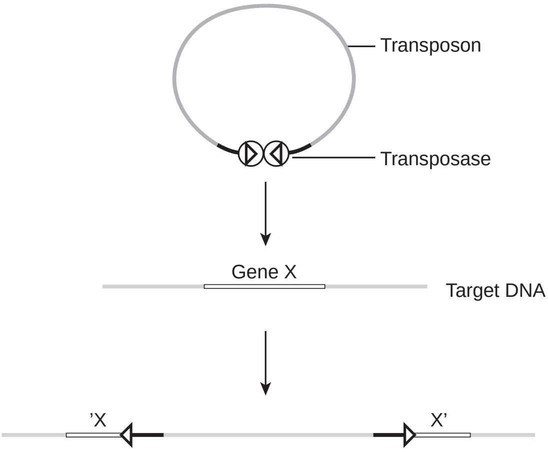


Fig. 1. Intermolecular transposition. The DNA transposons typically used for genetic analysis experiments are excised in a transposase-catalyzed fashion from their original genomic location. Pictured here are the next steps in transposition. The excised transposon complexed with transposase binds to target DNA (Gene X), and the transposase catalyzes integration of the transposon into Gene X thus generating ′X and X′ sequences. The transposase is presented as a circle. The specific end DNA sequences of the transposon are presented as open triangles.

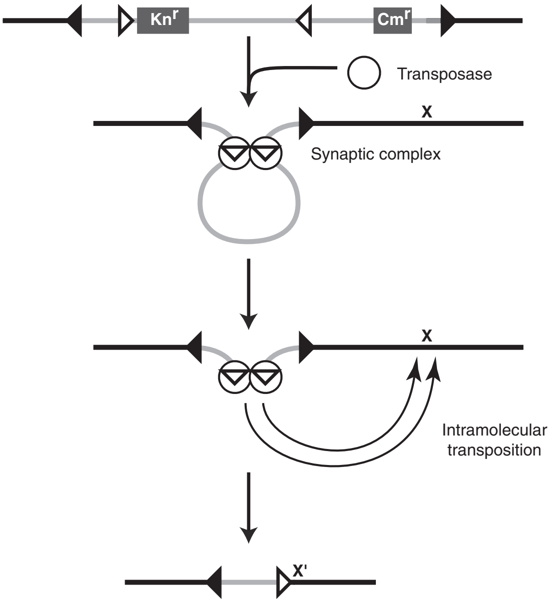


Fig. 2. Intramolecular transposition and adjacent deletion formation. A composite transposon is used for adjacent deletion formation. The composite transposon is made up of two transposable elements both defined by one open triangle and one closed triangle. Insertion events are first generated using intermolecular transposition of two closed triangles (not shown). An open triangle—specific transposase is synthesized, binds to open triangle ends, forms a synaptic complex, and then catalyzes intramolecular transposition to a site thousands of base pairs away, thus generating a deletion. This technology is described in more detail in Ref. *12*.

we shall describe deletion strategies that either use a random, transposition-based technology in which a composite transposon catalyzes inside-out intramolecular transposition (**Fig. 2**) or a transposon-to-transposon excision methodology between two insertions generated previously by standard transposition methodology (**Fig. 3**).

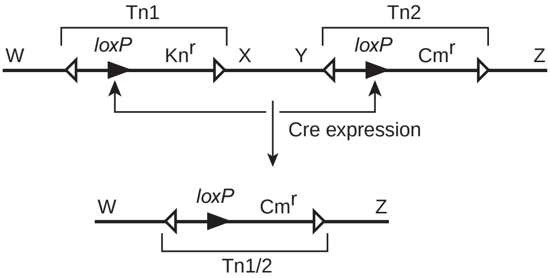


Fig. 3. Site-to-site deletion through Cre-catalyzed excision of DNA defined by two transposon inserts. Two Tn*5*-like inserts are separately generated through the electroporation of premade transposition complexes. Both transposons carry *lox*P sites (filled-in triangles), but one encodes kanamycin resistance (Kn^r) and the other encodes chloramphenicol resistance (Cm^r). Cre expression catalyzes the excision of DNA between the two *lox*P sites. See Ref. *19* for a more detailed description of the technology.

2. General Requirements

Hundreds of transposons have been identified and studied to some extent *(1)* so it would seem that a very large number of tools are possible. The transposition systems that are discussed in later chapters in this volume or related literature include Tn*3*, Tn*5*, Tn*7*, Tn*10*, Tn*4001*, and *mariner*. Although historical accidental choices certainly played a role in choosing these systems, the choice of transposon tools are restricted to ones that are well-enough studied so that we know that they can be made to fulfill the following requirements. First, the element must manifest a sufficiently high frequency of transposition through the desired protocol so that it is possible to achieve saturated mutagenesis (every gene hit at least once) in the organism's genome. Second, the targets chosen by a given transposition system should be sufficiently random so that any gene can suffer an insertion within the given procedure. It should be noted that all transposons likely manifest some degree of target sequence bias. Nonetheless, several of the transposition systems that are used as tools manifest a reasonable approximation of target randomness. Third, the transposition products should be genetically stable. This last criterion is typically achieved by not having the transposon-specific transposase synthesized in the target cells subsequent to the planned transposition event. Once these general requirements are met, the element of choice needs to be compatible with the transposition strategy used to generate the knockout libraries. We will describe below several different transposition strategies. Finally, the transposon of choice needs to contain the desired genetic markers demanded by the particular strategy. The most universal marker needed is an appropriate antibiotic resistance marker that will allow the selection of the desired transposition events in the particular host cell.

3. Transposon Structure

In general, natural transposons have the following basic structure (**Fig. 4**). They are defined by short (typically less than 50 bp), transposon-specific terminal DNA sequences. In many cases, these terminal sequences are inverted versions of the same or closely related sequences. The specific terminal inverted repeat sequences are key components of all the transposons that we shall use. Natural transposons also contain a gene encoding the transposon-specific transposase. The transposase binds specifically to the terminal inverted repeat sequences, forms a transposase-DNA synaptic complex, and, in the presence of Mg^{2+}, catalyzes the transposition events. Because of our need to generate genetically stable transposon inserts, the gene for transposase synthesis has been deleted from all of the constructs used in our studies. Instead, the transposase is encoded by a gene located outside of the transposon structure and is lost after transposition or else the transposase is provided biochemically. All transposons used in these studies encode antibiotic resistance in order to allow for the biological selection of the desired genetic events. Finally, transposons can be constructed to contain DNA sequences encoding other desired functions such as primer binding sites, T7 RNA polymerase promoters, site-specific recombination sites, genes encoding reporter functions, and so forth. In fact, transposons can carry any desired sequence as long as the

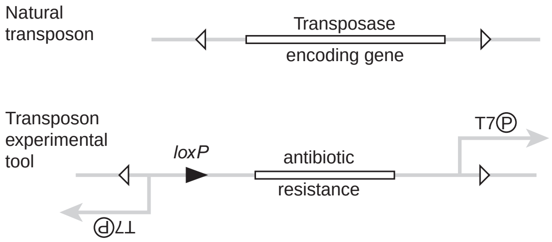


Fig. 4. Transposon structure. A natural DNA transposon has three components. The transposon ends are defined by two short (<50 bp), terminal, specific DNA sequences that typically are inverted versions of each other (open triangles). The transposon also encodes a transposase protein that catalyzes transposition of DNA sequences defined by the inverted terminal DNA sequences. Not shown are other genes that may be carried on the transposon. These other genes encode products that typically play no role in the transposition mechanism (i.e., antibiotic-resistance genes). By supplying the transposase exogenously, the transposon can be simplified as an experimental tool. In this case, the terminal transposase recognition sequences bracket DNA that contains the desired sequences. For example, the transposon can be constructed to contain an appropriate antibiotic resistance gene, outward-facing T7 promoters, and a *lox*P site.

length of DNA between the terminal inverted repeats is not so long (typically over several thousand base pairs) as to impair transposition.

4. Transposition Strategy

There is extensive literature that describes the use of transposons as genetic tools utilizing *in vivo* technologies; for instance, see the review by Berg et al. (*2*). These technologies utilize plasmid transformation or conjugation, or phage infection as a means for introducing the transposon into the target organism. The first adaptation was the use of plasmids or phages that were "suicide vectors." For suicide vectors, the phage genome or plasmid cannot be stably inherited by the target organism under the desired experimental conditions. The second property of suicide vectors is that the transposon-specific transposase is encoded by a gene that is contained on the phage or plasmid but outside of the transposon itself. Thus, after transposition and loss of the suicide vector, no transposase encoding sequence would be present and the transposition product would be genetically stable (**Fig. 5**). Systems that have utilized this type of *in vivo* strategy include the following examples: the signature-tagged mutagenesis (STM) Tn*5* system (*3*), the Tn*4001*-based individual knock-out system (*4*), the Tn*10*-based individual knock-out system (*5*), the *mariner*-based individual knock-out system (*6–9*), and the Tn*5*-based system for distinguishing cytoplasmic versus membrane proteins (*10, 11*). Finally, a modified version of a suicide vector strategy was used in the Tn*5*-based adjacent deletion technology (*12*).

A major accomplishment in transposition research was the development of *in vitro* transposition systems for a select group of transposons. The goal of this biochemical work was to enable research into the molecular basis of transposition, but the resulting

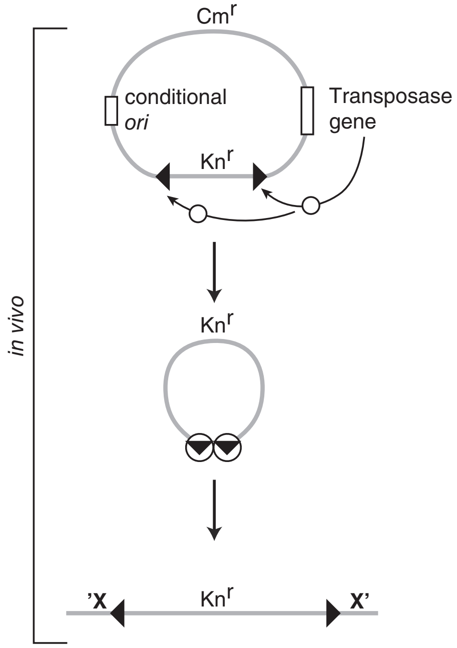
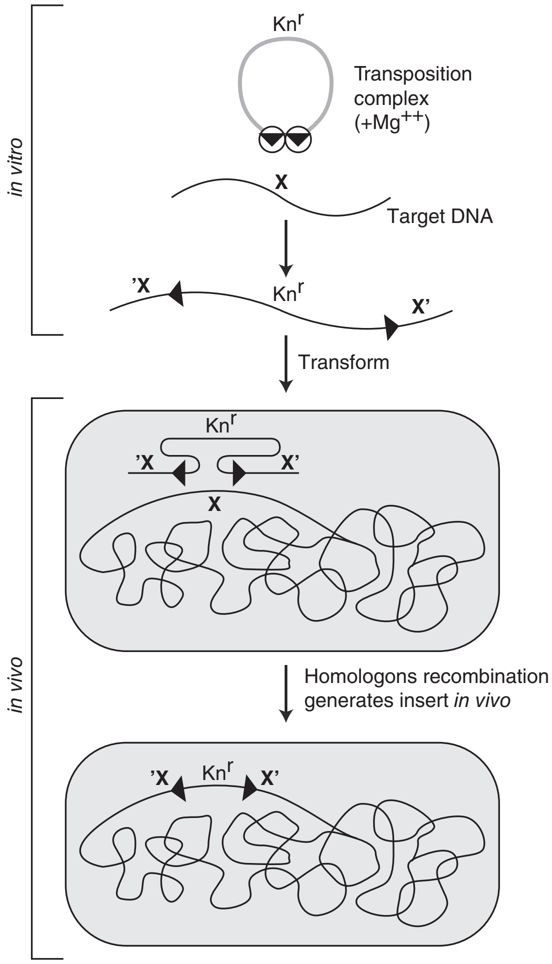


Fig. 5. Transposition mediated by a suicide vector system. The purpose of suicide vector systems is to allow *in vivo* transposition that results in genetically stable products; no subsequent transposition occurs after the initial insertion because no transposase is available. In this case, a suicide plasmid is utilized. The plasmid carries the transposon of choice (defined by solid triangles on either side of a Kn^r gene), an origin of replication (ori) that is unable to function in the chosen conditions, and a gene encoding the transposase that is located outside of the transposon. After plasmid introduction into the cell, the transposase (shown as open circles) is synthesized, and the transposase catalyzes transposition into the chromosome DNA and destruction of the plasmid (by formation of double-strand breaks). Because the plasmid is destroyed, no further synthesis of transposase occurs and no further transposition can occur. Similar phage-based suicide transposition systems have also been used.

technologies were soon adopted by investigators interested in applied uses of transposons such as the identification of essential genes. The general protocol involves *in vitro* transposition into target DNA followed by transformation of the DNA products into the target cells selecting for the presence of the transposon (**Fig. 6**). By this means, the transposon knockout strategy could be extended to organisms lacking a suitable *in vivo* suicide vector system (or allowed such a requirement to simply be bypassed). The critical requirement is that an efficient DNA transformation system must exist. Examples of the use of this *in vitro* technology can be found in the work by Akerley et al. ***(13)*** and Wong and Akerley ***(14)*** using the *mariner* transposition system, Kumar et al. ***(15)*** and Kumar ***(16)*** utilizing both the Tn*3* and Tn*7* systems, and Kang et al. ***(17)*** utilizing the Tn*5* system.

A system that combines both *in vitro* and *in vivo* manipulations involves the formation of transposon DNA—transposase complexes *in vitro* followed by electroporation of the transposition complexes (sometimes referred to as transposome or transpososome complexes) into the target cells ***(18)*** (**Fig. 7**). The *in vitro*—generated transposition complexes are catalytically activated when they encounter the intracellular Mg^{2+} leading to the random incorporation of the transposon into the cell's genome. This technology also bypasses any need for *in vivo* suicide vector strategies. The studies that have used this technology are described in Refs. ***12*** and ***19–22***.



in vitro
Kn^r
Transposition complex (+Mg^{++})
X
Target DNA
'X
Kn^r
X'
Transform
in vivo
Kn^r
'X
X'
X
Homologons recombination generates insert *in vivo*
'X
Kn^r
X'

Fig. 6. Use of *in vitro* transposition systems. *In vitro* transposition systems have been developed for some transposons. These *in vitro* systems allow the pictured transposition technology in which *in vitro* transposition is performed using purified target DNA and then the resulting transposition products are introduced into cells and incorporated into the cell's genome through homologous recombination.

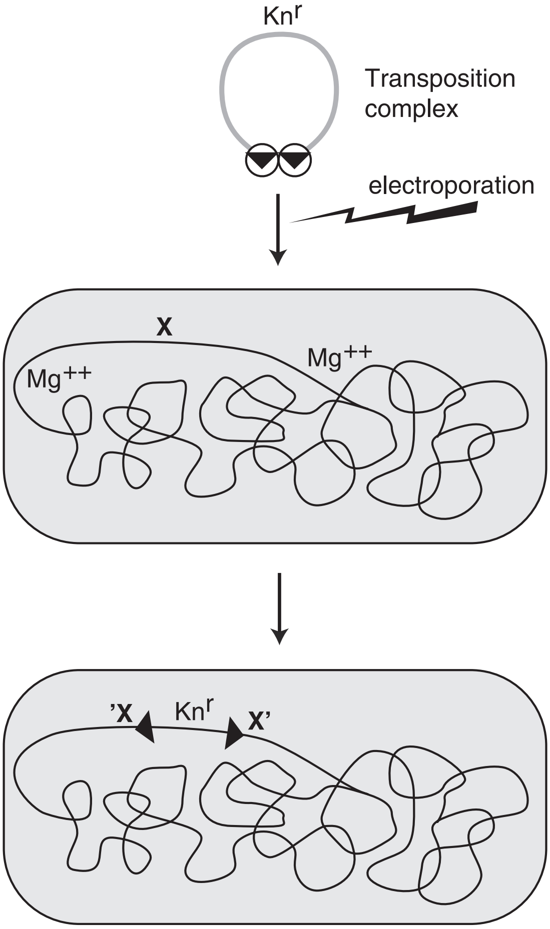


Fig. 7. Electroporation of preformed transposition complexes. Tn*5* transposase-transposon complexes give rise to transposition events after electroporation into a wide variety of target cells *(18)*.

5. Mapping/Detection Strategies

All of the one-by-one insertion mutation strategies described in this text are based on the proposal that pools (or libraries) of insertion mutants can be followed by various high-throughput techniques to determine how the individual mutants fare in competition with their peers found in the pool. This at first seems like a formidable challenge, but it has been achieved using a variety of technologies as described below.

As a first approach, a number of investigators have addressed the above challenge by first isolating individual transposon insertion mutants as colonies and then utilizing DNA sequence analysis of polymerase chain reaction (PCR)-amplified transposon-target junctions to define the gene location of each insert *(4, 8–11, 15, 16)* (**Fig. 8**). The sequenced inserts define nonessential genes. Once the PCR-amplified junction sequences are available, the ability of specific mutants to grow in pooled cultures under defined

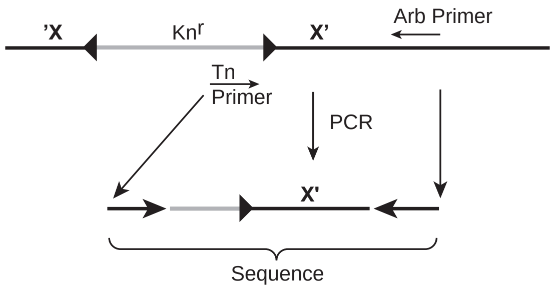


Fig. 8. Defining transposon inserts by sequence analysis of transposon—target DNA boundary sequences. A uniquely oriented transposon-specific primer (Tn primer) is coupled with an arbitrary (Arb) primer to PCR-amplify one of the transposon-target boundaries, which is subsequently sequenced in order to identify target DNA immediately adjacent to the transposon end sequence (filled triangle).

suboptimal conditions can be ascertained by nucleic acid hybridization analysis of the transposon-target joints.

A second approach is called "footprinting" (Fig. 9). The chromosome is divided into *in silico* segments whose lengths can be easily amplified by PCR. Primers are designed for each segment's ends. The inserts (in a much larger pool) are found within the defined segments by using a number of PCR reactions. Each PCR reaction is defined by a segment-specific primer and a transposon-specific primer. Typically, this experiment is used to define the end result of transposition plus outgrowth, but in some cases two PCR reactions are performed; one prior to growth (thereby defining the distribution of inserts in the inoculum) and one after outgrowth (thereby defining the viable mutants) (S. Gerdes, personal communication). This latter approach has the advantage of ruling

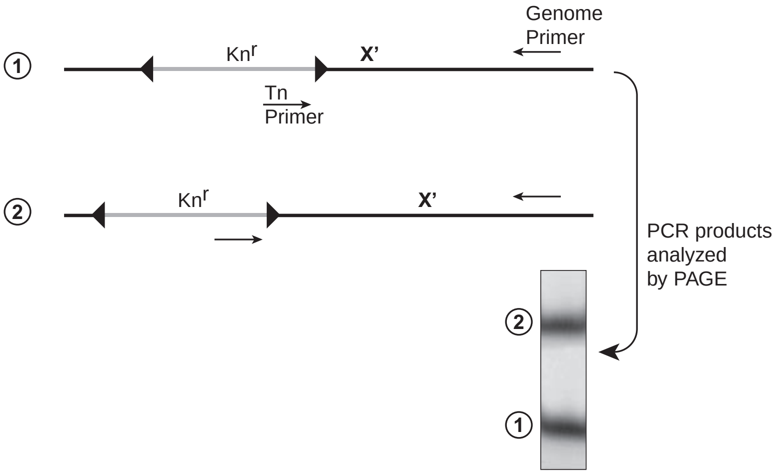


Fig. 9. Transposon footprinting. A large collection of transposon inserts are generated and pooled. The inserts in a general region are identified by performing a PCR reaction using two primers: a uniquely oriented transposon (Tn)-based primer and a primer corresponding with a given genomic site. All the transposon inserts in a given region can be identified by polyacrylamide gel electrophoresis (PAGE) of pooled PCR products.

out false-negatives; that is, the incorrect identification of genes as essential that merely do not serve as transposition targets for trivial reasons. Examples of footprinting can be found in Refs. *13, 14, 20,* and *21*.

Another technique that allows the analysis of mutant pools (and the mapping of mutant locations) involves microarray analyses *(5–7, 22)* (**Fig. 10**). The transposon used for generating the inserts carries T7 promoters facing out from both transposon ends. The DNA from a pool of inserts is extracted, cleaved with a restriction enzyme, and then used to program the synthesis of labeled RNA that is interrogated by hybridization to a microarray. In some applications *(5–7)* of this technology, the promoter-containing fragments are amplified by PCR and a cDNA copy of the RNA is generated.

Winterberg et al. *(22)* used high-density, whole-genome, custom-made oligonucleotide arrays from NimbleGen Systems, Inc. (Madison, WI). They were able to track the growth of individual inserts without resorting to promoter fragment amplification,

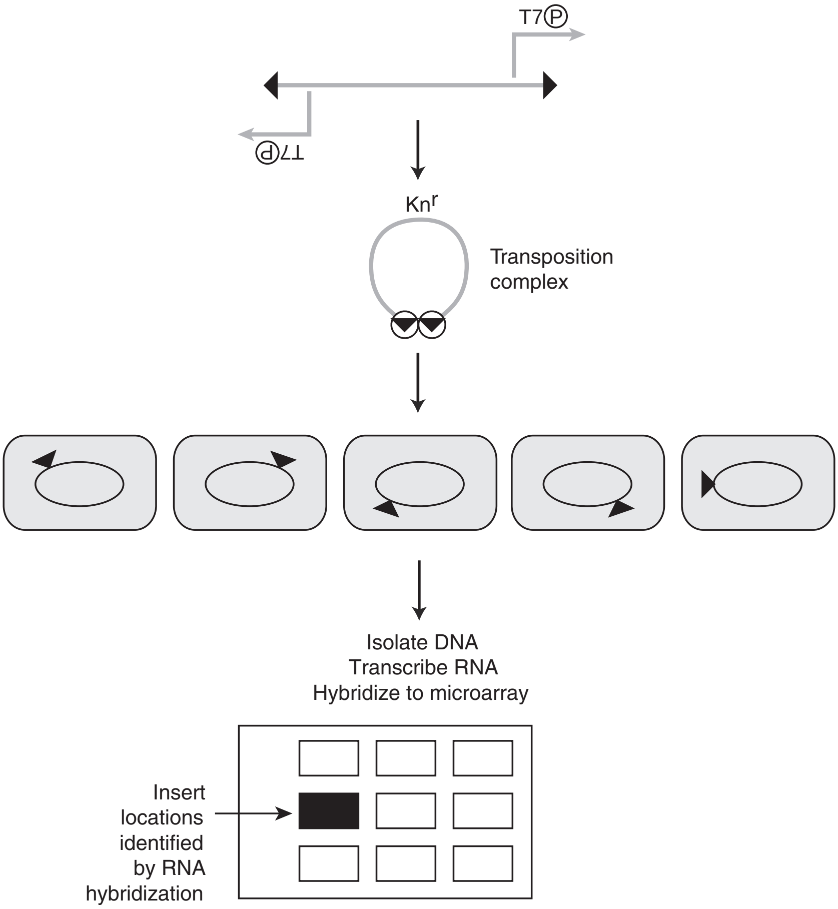


Fig. 10. Identification of pooled transposon insert locations through microarray analysis of transposon boundary transcripts. A transposon containing two outward-facing T7 promoters is used to generate a pool of inserts. The DNA from the transposition products is isolated, cleaved with a restriction enzyme, and used as templates for T7 RNA polymerase-catalyzed RNA synthesis. The location of the inserts is determined by microarray hybridization of the resulting RNA.

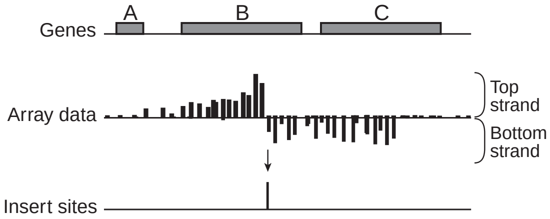


Fig. 11. Mapping inserts to within 50 bp with high-density microarrays. The experiment described in **Figure 10** is performed with high-density microarrays from NimbleGen Systems, Inc., in which 24 nt oligonucleotides correspond with each strand of DNA and are spaced at approximately 50-nt intervals throughout the entire genome. The resulting data can be analyzed to identify the site of insertion for each of ~100 inserts to within about 50 nt. This figure is similar to a figure presented in Ref. *22*.

they were not limited to studying inserts within known open reading frames (ORFs), and they were also able to map the insert locations to within ~50 bp. The method for mapping inserts is diagrammatically presented in **Figure 11**.

The last technique that we shall review is the oldest: STM *(3)* (**Fig. 12**). In this technology, each transposon in the mutagenesis collection is constructed to contain one

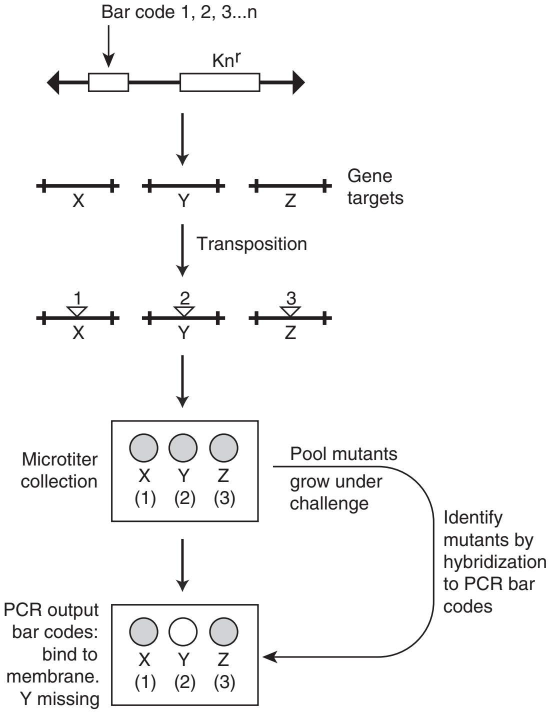

Bar code 1, 2, 3...n
Kn^r
X
Y
Z
Gene targets
Transposition
1
2
3
X
Y
Z
Microtiter collection
X (1)
Y (2)
Z (3)
Pool mutants grow under challenge
Identify mutants by hybridization to PCR bar codes
PCR output bar codes: bind to membrane. Y missing
X (1)
Y (2)
Z (3)

Fig. 12. Tracking inserts using signature-tagged mutagenesis (STM). Inserts are generated using transposons that carry a variety of 20-bp random sequences (bar codes). The resulting insert mutants are individually distributed into microtiter wells. The mutants are mixed to form a pool that is challenged to grow under some defined condition. The input and output pool DNAs are subjected to PCR amplification of the bar codes, which are hybridized to membrane representations of the individual microtiter arrays of the mutants leading to the discovery of which mutants did and did not grow under the challenging conditions.

of a multiplicity of 20-mer random sequences. Thus, upon mutagenesis, each mutant is uniquely defined by the specific 20-mer. Ninety-six separate mutant DNAs (and thus 96 different 20-mers) are arrayed on a hybridization detection membrane. Mutants are pooled and interrogated before (input) and after (output) challenging outgrowth (or colonization) by labeling PCR-amplified identifying 20-mers found in the mixed culture DNAs and hybridizing them to membranes imprinted with 96 mutant DNAs each. Missing mutants are readily identified by a failure to detect a hybridization signal between input and output samples. The transposon insertion contained in the missing mutant can be characterized by PCR and sequencing of the transposon DNA junctions.

6. Identifying Essential Functions: Serial Deletion Generation

It should be possible to define a minimal genome sequence, that is, a gene complement that encodes all essential functions, by attempting to shrink the genome size through repetitive deletion generation. Transposons have been used in this process through two entirely different technologies: one is random in nature and the other is semidirected. Both technologies represent work in progress because neither is likely near defining the minimal required genome.

The random transposon-mediated deletion approach is depicted in **Figure 2** *(12)*. This technology utilizes well-known aspects of transposon biology. A composite transposon is used in which the transposon is constructed from two transposable elements each defined by the same two different terminal DNA sequences and each inverted relative to each other. Thus, the composite transposon contains two "inside" terminal DNA sequences and two "outside" terminal sequences. A second property of the technology is that it uses both intermolecular transposition and intramolecular transposition. The intermolecular transposition, involving the "outside" terminal sequences, is accomplished via electroporation of preformed transposase-transposon complexes to generate random inserts into the bacterial chromosome. Intramolecular transposition from the "inside" terminal sequences utilizes *in vivo* transposition catalyzed by "inside"-specific transposases to generate adjacent deletions that extend out from the transposon insert to a site on the adjacent chromosome (typically ~10,000 bp away). This random transposon-mediated deletion approach has been repeated 47 times to generate a MG1655 derivative lacking ~14% of its original genome, obviously a long way from achieving a minimal genome (J. Apodaca, personal communication). Already, however, 55 genes that were previously reported as essential have been deleted with limited physiologic effects (an elongated lag time in rich medium).

The above technology can be modified to allow the deleted material at each cycle to be maintained as a plasmid. This would allow a conditional assessment of the role of deleted material by analyzing the cells before and after loss of the deletion-generated plasmid *(12)*.

A semidirected deletion approach, using transposons as tools, has also been reported *(19, 23)* (**Fig. 3**). Tn*5* transposons modified to contain one of two different antibiotic-resistance genes and a *lox*P site were randomly inserted into the *Escherichia coli* genome utilizing electroporation of premade transposition complexes. Strains were generated with two different inserts and then Cre site-specific recombinase was used to excise the genomic material between the two sites. Survival and growth of the resulting mutants indicated that no essential genes were removed. Some individual deletions were combined to achieve a reduction of more than 300 kbp. Again, this work is a long way from achieving a minimalized chromosome but the technology is available and in use.

6. Conclusion

Transposons are important genetic tools for performing genetic analyses. The current chapter and many of the chapters in this text describe how transposons can be used to define essential genes. These technologies are based on two strategies: a one-by-one knockout strategy that identifies which genes can be interrupted and still allow growth (and by subtraction, which genes are not found to suffer mutations and are thus likely essential), and deletion-based approaches that remove several genes at once. The deletion approach could lead to an identification of the minimal required bacterial genome for growth under specified conditions.

References

1. Chandler, M., and Mahillon, J. (2002) Insertion sequences revisited. In: Craig, N., Craigie, R., Gellert, M., and Lambowitz, A. M. eds. *Mobile DNA II*. Washington, DC: ASM Press, pp. 305–366.
2. Berg, C. M., Berg, D. E., and Groisman, E. (1989) Transposable elements and the genetic engineering of bacteria. In: Berg, D. E., and Howe, M. M., eds. *Mobile DNA*. Washington, DC: ASM Press, pp. 879–925.
3. Hensel, M., Shea, J. E., Gleeson, C., Jones, M. D., Dalton, E., and Holden, D. W. (1995) Simultaneous identification of bacterial virulence genes by negative selection. *Science* **269**, 400–403.
4. Hutchison, C. A. III, Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., et al. (1999) Global transposon mutagenesis and a minimal Mycoplasma genome. *Science* **286**, 2165–2169.
5. Badarinarayana, V., Estep, P. W. III, Shendure, J., Edwards, J., Tavazoie, S., Lam, F., and Church, G. M. (2001) Selection analyses of insertional mutants using subgenic-resolution arrays. *Nat. Biotech.* **19**, 1060–1065.
6. Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2001) Comprehensive identification of conditionally essential genes in mycobacteria. *Proc. Nat. Acad. Sci. U.S.A.* **98**, 12712–12717.
7. Murry, J. P., Sassetti, C. M., Lane, J. M., Xie, Z., and Rubin, E. J. (2007) Transposon site hybridization (TraSH) in *Mycobacterium tuberculosis*. This volume, Chapter 4.
8. Liberati, N. T., Urbach, J. M., Miyata, S., Lee, D. G., Drenkard, E., Wu, G., et al. (2006) An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. *Proc. Nat. Acad. Sci. U.S.A.* **103**, 2833–2838.
9. Liberati, N. T., Urbach, J. M., Holmes, T. K., Wu, G., and Ausubel, F. M. (2006) Comparing insertion libraries in two *Pseudomonas aeruginosa* strains to assess gene essentiality. This volume, Chapter 10.
10. Jacobs, M. A., Alwood, A., Thaipisuttikul, I., Spencer, D., Haugen, E., Ernst, S., et al. (2003) Comprehensive transposon mutant library of *Pseudomonas aeruginosa. Proc. Nat. Acad. Sci U.S.A.* **100**, 14339–14344.

11. Jacobs, M. A. (2006) How to: make a defined near-saturation mutant library. Case 1: *Pseudomonas aeruginosa* PAO1. This volume, Chapter 9.
12. Goryshin, I. Y., Naumann, T. A., Apodaca, J., and Reznikoff, W. S. (2003) Chromosomal deletion formation system based on Tn5 double transposition: use for making minimal genomes and essential gene analysis. *Genome Res.* **13**, 644–653.
13. Akerley, B. J., Rubin, E. J., Camilli, A., Lampe, D. J., Robertson, H. M., and Mekalanos, J. J. (1998) Systematic identification of essential genes by *in vitro mariner* mutagenesis. *Proc. Nat. Acad. Sci. U.S.A.* **95**, 8927–8932.
14. Wong, S. M. S., and Akerley, B. J. (2006) Identification and analysis of essential genes in *Haemophilus influenzae*. This volume, Chapter 3.
15. Ross-Macdonald, P., Coelho, P. S., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., et al. (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption**.** *Nature* **402**, 413–418.
16. Kumar, A. (2006) Multipurpose transposon insertion libraries for large-scale analysis of gene function in yeast. This volume, Chapter 8.
17. Kang, Y., Durfee, T., Glasner, J. D., Qiu, Y., Firsch, D., Winterberg, K., and Blattner, F. R. (2004) Systematic mutagenesis of the *Escherichia coli* genome. *J. Bacteriol.* **186**, 4921–4930.
18. Goryshin, I. Y., Jendrisak, J., Hoffman, L., Meis, R., and Reznikoff, W. S. (2000) Insertional transposon mutagenesis by electroporation of released Tn5 transposition complexe**s.** *Nat. Biotech.* **18**, 97–100.
19. Yu, B. J., Sung, B. H., Koob, M. D., Lee, C. H., Lee, J. H., Lee, W. S., et al. (2002) Minimization of the *Escherichia coli* genome using a Tn*5*-targeted Cre/*loxP* excision system**.** *Nat. Biotech.* **20**, 1018–1023.
20. Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balázsi, G., Rayasz, E., Daugherty, M.D., et al. (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684.
21. Scholle, M. D., and Gerdes, S. Y. (2006) Whole-genome detection of conditionally essential and dispensable genes in *E. coli* via genetic footprinting. This volume, Chapter 6.
22. Winterberg, K. M., Luecke, J., Bruegl, A. S., and Reznikoff, W. S. (2005) Phenotypic screening of *Escherichia coli* K-12 Tn5 insertion libraries using whole genome oligonucleotide microarrays. *Appl. Environ. Microbiol.* **71**, 451–459.
23. Yu, B. J., and Kim, S. C. (2006) Minimization of the *Escherichia coli* genome using the Tn*5*-targeted Cre/*loxP* excision system. This volume, Chapter 17.

3

Identification and Analysis of Essential Genes in *Haemophilus influenzae*

Sandy M.S. Wong and Brian J. Akerley

Summary

The human respiratory pathogen *Haemophilus influenzae*, a Gram-negative bacterium, is the first free-living organism to have its complete genome sequenced, providing the opportunity to apply genomic-scale approaches to study gene function. This chapter provides an overview of a highly efficient, *in vitro mariner* transposon–based method that exploits the natural transformation feature of this organism for the identification of essential genes. In addition, we describe strategies for conditional expression systems that would facilitate further analysis of this class of genes. Finally, we outline a method based on the approach used in *H. influenzae* for identifying essential genes that can be applied to other bacteria that are not naturally transformable.

Key Words: conditional expression; essential genes; GAMBIT; *H. influenzae*; *mariner* transposon *in vitro* mutagenesis; mutagenic PCR; SCE jumping.

1. Introduction

The availability of complete bacterial genome sequences has ushered in an era of microbial functional genomics. This abundance of sequence information presents challenges and opportunities, such as the discovery of genes whose functions have not been identified. Comparative computational approaches have proved useful for addressing the roles of such genes, but experimental approaches are needed both to evaluate sequence-based hypotheses and to extend our knowledge to previously uncharacterized genes and biological functions. Essential genes represent an attractive category for functional analysis because they mediate primary cellular functions and are potential targets for antimicrobial agents. Numerous investigators have reported various global approaches for identifying genes of essential function under defined growth conditions in bacteria ***(1–18)***, some of which are discussed elsewhere in this book.

For the purpose of this discussion, essential genes are those that cannot be inactivated in otherwise wild-type bacteria without abrogating growth or survival in culture on rich medium, though essentiality is ultimately context-dependent. In this chapter, a

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

methodology is described that is based on genetic footprinting *(19)* and was applied to querying large numbers of genes for essential functions in *Haemophilus influenzae* (*20, 21*). Unlike stochastic genetic screens, this approach involves systematic mutagenesis of specific genomic regions allowing comprehensive analysis of the entire genome. An added benefit of this method is that it generates an ordered bank of mutants carrying transposon insertions in every nonessential gene, and these can be readily recovered for further characterization. Verification and further study of essential genes can be accomplished by construction of strains containing temperature-sensitive (TS) or conditionally expressed essential proteins. The procedures used in *H. influenzae* to generate these types of strains are discussed. Although developed for the analysis of essential genes, the conditional expression system is equally applicable to study of genes required for genetic stability of the bacterium, such as factors involved in DNA repair or protection from oxidative damage. We also present one method by which this general approach for identification of essential or conditionally essential genes can be extended to other bacteria.

2. Materials

1. Naturally transformable *H. influenzae* strain (e.g., Rd KW20).
2. Brain heart infusion (BHI) broth and BHI agar supplemented with 10 µg/mL nicotinamide adenine dinucleotide and 10 µg/mL hemin (sBHI).
3. Media containing kanamycin (Km) at 20 µg/mL and tetracycline (Tet) at 8 µg/mL for *H. influenzae*; gentamicin at 5 µg/mL and 150 µg/mL for *Escherichia coli* and *Pseudomonas aeruginosa*, respectively.
4. Oligonucleotide primers.
5. D(+)-Xylose (Sigma-Aldrich, St. Louis, MO).
6. Restriction enzymes.
7. T4 DNA polymerase (New England Biolabs [NEB], Beverly, MA).
8. T4 DNA ligase (NEB).
9. *Himar1* transposase.
10. *Taq* polymerase.
11. Pfu polymerase (Stratagene, La Jolla, CA).
12. T4 DNA ligase buffer (NEB).
13. NEB2 restriction buffer (NEB).
14. Acetylated BSA (NEB).
15. Transposition buffer: 10% glycerol v/v, 25 mM Hepes pH 7.9, 100 mM NaCl, 5 mM $MgCl_2$, 2 mM dithiothreitol (DTT), and 25 µg/mL acetylated BSA.
16. Deoxyribonucleotide triphosphate (dNTP).
17. QIAquick PCR Purification Kit (Qiagen, Valencia, CA).
18. Gel filtration cartridges (Edge BioSystems, Gaithersburg, MD).
19. Thermal cycler.
20. Electroporator.
21. Electrophoresis equipment.
22. *E. coli* strains S17-1, TOP10, SM10.
23. Luria-Bertani (LB) agar and broth.
24. *P. aeruginosa* strain PAO1.
25. Membrane filters (0.2-µm analytical test filter funnel; Fisher Scientific, Pittsburgh, PA).

3. Methods

The methods described below outline the use of GAMBIT (genomic analysis and mapping by *in vitro* transposition) to define essential regions of the *H. influenzae* chromosome (**Section 3.1**), SCE-jumping in *P. aeruginosa*, an approach adapted from GAMBIT for mutagenesis of targeted regions of the chromosome in bacteria that are not naturally transformable (**Section 3.2**), and two methods to functionally characterize individual essential genes in *H. influenzae* (**Section 3.3**). These methods include the use of a conditional expression system and mutagenic polymerase chain reaction (PCR)-generated TS mutations in *H. influenzae*.

Detailed methodology for purifying *mariner* transposase and for conducting *in vitro mariner* transposition reactions has been described previously *(22)*; however, an updated description of the basic steps of the mutagenesis procedure is outlined (**Section 3.1**). We also include in **Note 2** functional analysis of the minimal terminal inverted repeat sequences of a *Himar1* (*mariner*)-derived minitransposon. Note, however, that the *in vitro* transposition can be performed with other transposase/transposon combinations, provided that a repair step is included to allow uptake by naturally transformable bacteria. The focus of this chapter is on the application of these and other methods to the identification and study of essential genes.

3.1. GAMBIT in Haemophilus influenzae

Figure 1 outlines a general scheme for generating transposon insertion mutants in *H. influenzae* by *in vitro* transposon mutagenesis. The procedure takes advantage of *in vitro* transposition by a *mariner*-derived minitransposon and the natural transformability of *H. influenzae*, which can acquire DNA from its external environment and efficiently incorporate it into its genome. Incorporation of foreign sequences into the genome by homologous recombination requires a minimum of ~150 bp of homologous flanking sequence and the presence of a DNA uptake sequence (US) consisting of a highly conserved 9-bp core sequence within a 29-bp consensus sequence for optimal transformation efficiency *(23)*. The *in vitro* transposition reaction consists of target DNA (either a PCR product or chromosomal DNA), donor transposon DNA (such as the *mariner*-family transposon *Himar1*), and purified *Himar1* transposase. Construction of plasmid pENTUS carrying a *mariner*-derived transposon is described elsewhere *(21)*. Briefly, pENTUS carries a kanamycin (Km) resistance cassette and an *H. influenzae* uptake sequence flanked by *Himar1* inverted terminal repeat sequences. (Functional studies to assess the minimal *Himar1* repeat sequences necessary for efficient transposition are described in **Note 2** and **Fig. 6**.) Transposition of *Himar1* is very efficient, requiring only the transposase without any cofactors derived from host cells. This transposon shows little target site specificity, inserting at the dinucleotide TA in the target sequence. The minimal site specificity of the *mariner* transposase represented a particularly major breakthrough for genetic studies of *H. influenzae*, for which the primary transposon mutagenesis tool had previously been the Tn*916* transposon. For example, an *in silico* evaluation of Tn*916* insertion sites in the *H. influenzae* KW20 genome sequence yielded 167 potential target sites, and only 80 of these were in open reading frames *(24)*.

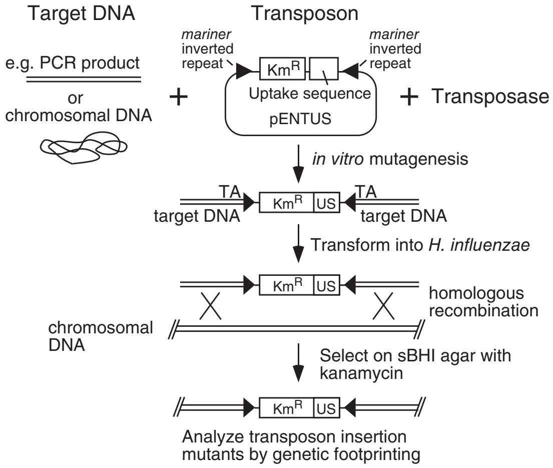


Fig. 1. *In vitro* transposon mutagenesis in *H. influenzae*. Schematic diagram illustrates overview of *in vitro* mutagenesis with a kanamycin-marked (Km) *mariner* transposon carried on a plasmid that does not replicate in *H. influenzae* containing a *H. influenzae* uptake sequence (US). The *mariner* transposon inserts at TA dinucleotides in the target sequence, resulting in a duplication of the TA dinucleotide flanking the insertion. Mutagenized DNA is introduced into *H. influenzae* by natural transformation. Transformants are selected on kanamycin-containing medium and analyzed by PCR for genetic footprinting analyses. (Adapted with permission from Ref. *21*. Copyright © 2002 National Academy of Sciences, U.S.A.)

The overall GAMBIT procedure is listed in the following steps below. Detailed descriptions of each step are described in **Section 3.1.1** through **Section 3.1.6**. **Figure 2** illustrates a detailed scheme of the GAMBIT procedure for identifying essential genes in *H. influenzae* under a specific growth condition.

1. PCR amplify target DNA of interest (e.g., 10 kb chromosomal region).
2. Perform *in vitro* mutagenesis with *mariner* transposon (marked with an antibiotic cassette, e.g., kanamycin) and transposase.
3. Repair single-stranded gaps introduced on either side of the transposon insertion by the transposase with T4 DNA polymerase and T4 DNA ligase.
4. The following reaction is then transformed into *H. influenzae* that is made naturally competent *(25)*. The incoming DNA is integrated into the genome by homologous recombination (usually ~500 bp of flanking sequences is sufficient for efficient recombination).
5. Transformants are selected on sBHI containing kanamycin.
6. Pooled kanamycin-resistant transformants are analyzed by PCR for genetic footprinting *(19)*.

3.1.1. Amplification of Target DNA

The *in vitro* transposition system allows high-density transposon mutagenesis of discrete subgenomic regions. For the genome analysis of *H. influenzae* essential genes, ~10-kb chromosomal regions, overlapping by ~5 kb and covering the entire genome (1830 bp), were systematically amplified by PCR. In these experiments, primers were designed using the MacVector sequence analysis program to have a 62°C theoretical melting temperature, 40% to 60% A+T composition, and absence of 3′ dimer formation.

1. To amplify target DNA for *in vitro* mutagenesis, combine 1/10th volume of 10X DNA Thermopol buffer (NEB), using Taq polymerase and Pfu polymerase at a 10 : 1 unit ratio (10 units per reaction), 100 pmol of primers, and ~20 ng of chromosomal DNA as template.
2. Amplify using the following cycling parameters: 30 cycles of amplification with 30 s denaturation at 95°C, 30 s annealing at 62°C, and 5 min extension at 68°C with 15 s added to the extension time at each cycle.
3. PCR products are purified using the QIAquick PCR Purification Kit.

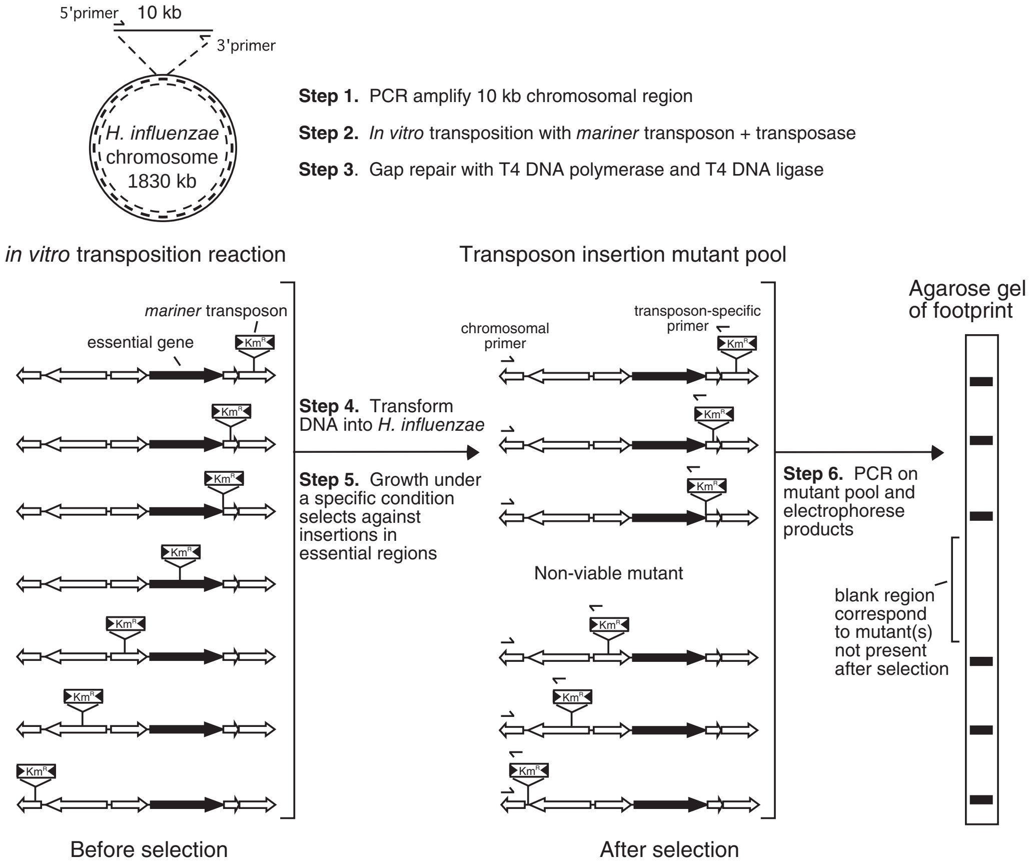

5'primer
10 kb
3'primer
H. influenzae
chromosome
1830 kb
Step 1. PCR amplify 10 kb chromosomal region
Step 2. *In vitro* transposition with *mariner* transposon + transposase
Step 3. Gap repair with T4 DNA polymerase and T4 DNA ligase
in vitro transposition reaction
Transposon insertion mutant pool
Agarose gel of footprint
mariner transposon
essential gene
Km^R
chromosomal primer
transposon-specific primer
Step 4. Transform DNA into *H. influenzae*
Step 5. Growth under a specific condition selects against insertions in essential regions
Step 6. PCR on mutant pool and electrophorese products
Non-viable mutant
blank region correspond to mutant(s) not present after selection
Before selection
After selection

Fig. 2. GAMBIT in *H. influenzae*. Chromosomal regions of interest are mutagenized *in vitro* with *mariner* transposon and *Himar1* transposase. Mutagenized DNA is transformed into *H. influenzae* and integrates into the genome by homologous recombination. After selection on medium containing kanamycin, the transposon insertion mutants are pooled and used as template for genetic footprinting by PCR with a chromosomal primer and a *mariner*-specific primer. Mutants that have sustained an insertion in an essential gene will drop out of the pool under the specific growth condition and will not be represented with a corresponding PCR product, giving rise to a blank region on an agarose gel. (Adapted with permission from Refs. ***20*** and ***21***. Copyright © 1998 and 2002 National Academy of Sciences, U.S.A.)

3.1.2. Transposition Reaction

1. For each transposition or control reaction, combine target DNA (up to 1 µg) and *mariner* transposon donor DNA (~100 to 500 ng or to a 5 : 1 target-to-donor molar ratio), and 20 µL of 2X Transposition buffer (20% glycerol v/v, 50 mM Hepes pH 7.9, 200 mM NaCl, 10 mM $MgCl_2$, 4 mM DTT, and 50 µg/mL acetylated BSA from NEB).
2. Adjust to a volume of 39 µL with distilled water.
3. Add 1 µL *Himar1* transposase to a final concentration of ~10 nM for a total of volume of 40 µL. Purification of transposase is described in detail elsewhere ***(22)***. In a control mock-transposition processed in parallel, add 1 µL of distilled water.
4. Incubate at 37°C for 1 to 6 h.
5. Purify the transposition reaction with either QIAquick spin column (Qiagen) (elute with 30 µL water, pH 7 to 8.5) or phenol/chloroform extraction followed by ethanol precipitation and resuspension in 30 µL water.

3.1.3. Repair Reaction with T4 DNA Polymerase

1. On ice, combine the following for each transposition reaction: T4 DNA polymerase (0.5 µL of 3000 U/mL from NEB), 4 µL 10X NEB buffer 2 (500 mM NaCl, 100 mM Tris-HCL, pH 7.9 at 25°C, 100 mM $MgCl_2$, and 10 mM DTT), 4 µL 2.5 mM dNTPs, and 1.5 µL distilled water for a total volume of 10 µL.
2. Gently mix the 10 µL T4 polymerase mixture with 30 µL of the purified transposition reaction or mock reaction described in **Section 3.1.2** for a total volume of 40 µL.
3. Incubate at room temperature for 20 min.
4. Terminate reaction by heating at 75°C for 15 min.

3.1.4. Repair Reaction with T4 DNA Ligase

1. To each 40 µL heat-killed T4 DNA polymerase reaction, add the following mixture: 1 µL T4 DNA ligase (40,000 U/mL from NEB), 12 µL 10X T4 DNA ligase buffer (500 mM Tris-HCL, pH 7.5 at 25°C, 100 mM $MgCl_2$, 100 mM DTT, 10 mM ATP, and 250 µg/mL BSA), and 67 µL of distilled water for a total volume of 120 µL.
2. Incubate at 16°C from 4 h to overnight.
3. Terminate reaction at 75°C for 15 min.
4. Purify ligation reaction with QIAquick spin columns (Qiagen).
5. Elute with 30 µL to 50 µL water, pH 7 to 8.5. Alternatively, purify reaction products with gel filtration cartridges (Edge BioSystems).

3.1.5. H. influenzae *Transformation and Postselection*

1. Use 15 µL to 25 µL of the repaired T4 DNA ligase reaction from above for *H. influenzae* transformation. Include a mock transformation in parallel containing cells with no added DNA and, as a transformation standard, transform cells with a known concentration of the Km resistance gene flanked by *H. influenzae* sequences corresponding with a nonessential gene.
2. Select transformants for colony formation on sBHI agar containing Km at 35°C. Mutants that have sustained an insertion in an essential gene that is required for optimal growth under the specific *in vitro* conditions become nonviable and will be absent after selection. With a highly competent cell preparation, approximately 100 to 1000 mutants are typically obtained for each region.

3.1.6. Genetic Footprinting

1. The Km-resistant transformants are pooled in 10 mM Tris buffer (pH 8.0) and diluted to an OD_{600} of 0.1, and 1 µL is used immediately as template in PCR for genetic footprinting. Alternatively, genomic DNA from the mutant pool can be purified and used as template in PCR, although we have found that purification is not necessary. PCR conditions are identical to those described in **Section 3.1.1**. Each primer specific to chromosomal sequences (designed in **Section 3.1.1**), in combination with a primer specific to *mariner* transposon sequences, is used to map transposon insertions within each pool of mutants. To allow mapping independent of the orientation of transposon insertions, we designed a primer, Marout (5′CCGGGGACTTATCAGCCAACC), that binds to both inverted repeats and primes outward from the transposon.
2. The PCR reaction is analyzed by agarose gel electrophoresis to generate a genetic footprint. Nonviable mutants that have sustained a transposon insertion in an essential gene will not be represented by a corresponding PCR product, producing a blank region or "window" on the gel. It is important to note, however, that many essential genes can tolerate insertions in the extreme 3′ end *(**20, 26**)*. Conversely, nonessential genes will contain *mariner* insertions in nearly every TA dinucleotide, resulting in a "ladder" of bands corresponding with the distance between each insertion and the position of the chromosome specific primer (**Fig. 2**).

3.2. A Variation of the H. influenzae *GAMBIT Procedure*

The application of the GAMBIT procedure has facilitated functional genomic analyses of essential genes in *H. influenzae* by exploiting the efficiency of the *in vitro mariner* transposition system. The *Himar1 in vitro* transposition system has been successfully adapted for other naturally transformable bacteria that can take up exogenous naked DNA for chromosomal integration such as *Streptococcus pneumoniae (**20**)* and *Campylobacter jejuni (**27**)*. As more bacterial genomes are sequenced, it would be beneficial to establish an analogous mutagenesis system for functional gene analyses in bacteria that are not naturally transformable. *Pseudomonas aeruginosa*, a major opportunistic human pathogen, has one of the largest bacterial genomes (6.3 Mbp) sequenced to date with as many as 50% of the total open reading frames composed of genes of unknown function *(**28**)*. An adaptation of the *H. influenzae* GAMBIT procedure for bacteria that are not naturally transformable, such as *P. aeruginosa*, would require features that allow efficient delivery of mutations to targeted regions of the chromosome. Because conjugation has proved to be a highly efficient method of introducing DNA into *P. aeruginosa*, this feature was incorporated into a strategy to develop an allelic exchange system, termed "SCE jumping," that delivers transposon insertions to discrete regions of the chromosome *(**29, 30**)*. Two crucial features contributing to the success of the SCE jumping allelic exchange method are the use of the *mariner* transposon for highly efficient *in vitro* transposition and expression of the yeast endonuclease I-*Sce*I in *P. aeruginosa*. This enzyme recognizes an 18-bp recognition site (5′-TAGGGATAACAG GGTAAT-3′) that is absent in bacterial genomes sequenced to date. The I-*Sce*I sites are designed to flank *P. aeruginosa* cloned DNA containing a *mariner* transposon–encoded resistance determinant (e.g., gentamicin) carried on an allelic exchange vector that is mobilized from an *E. coli* donor via conjugation (**Fig. 3**). Presence of I-*Sce*I in *P. aeruginosa* catalyzes the cleavage of the double-strand plasmid DNA at the

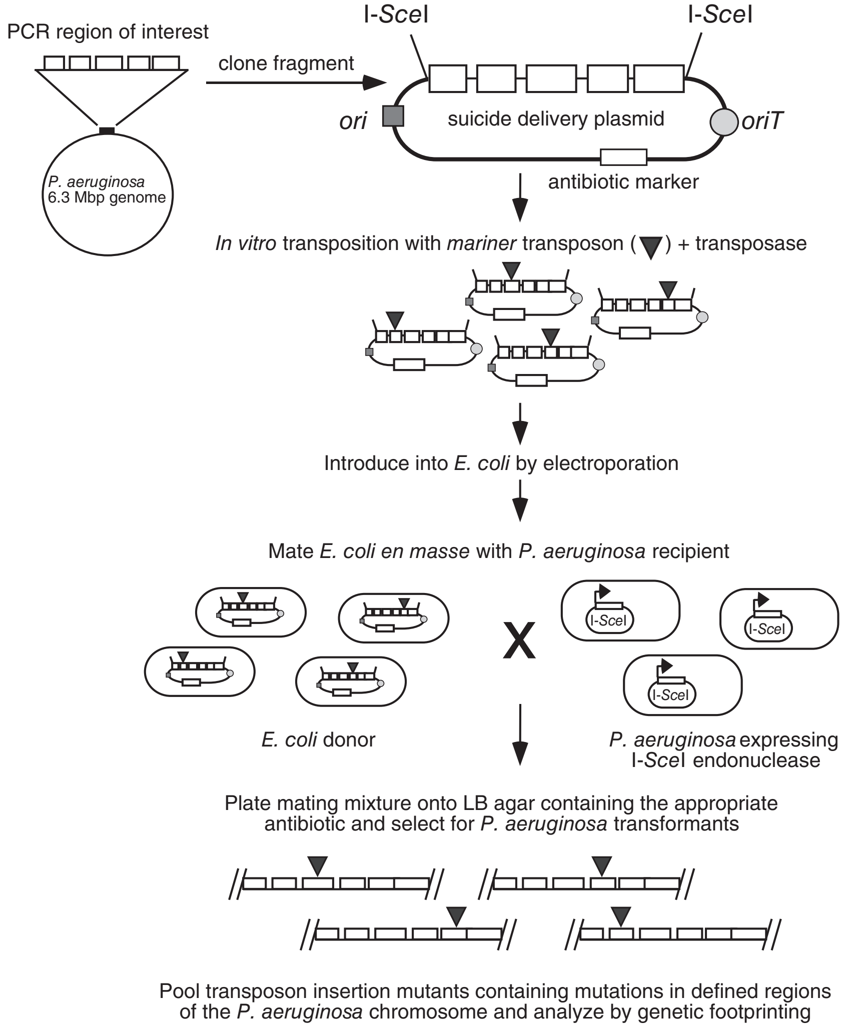

PCR region of interest
clone fragment
I-*Sce*I
I-*Sce*I
ori
suicide delivery plasmid
oriT
P. aeruginosa 6.3 Mbp genome
antibiotic marker
In vitro transposition with *mariner* transposon (▼) + transposase
Introduce into *E. coli* by electroporation
Mate *E. coli en masse* with *P. aeruginosa* recipient
X
I-*Sce*I
I-*Sce*I
I-*Sce*I
E. coli donor
P. aeruginosa expressing I-*Sce*I endonuclease
Plate mating mixture onto LB agar containing the appropriate antibiotic and select for *P. aeruginosa* transformants
Pool transposon insertion mutants containing mutations in defined regions of the *P. aeruginosa* chromosome and analyze by genetic footprinting

Fig. 3. SCE jumping in *P. aeruginosa*: an approach for high-density mutagenesis of a targeted chromosomal region. A PCR product containing the region of interest is cloned between I-*Sce*I sites contained on a plasmid that does not replicate in *P. aeruginosa* (e.g., plasmids containing ColE1 origin of replication) and mutagenized *in vitro*. The mutant plasmid pool is introduced into *E. coli*, and recombinants are selected for drug resistance encoded by the transposon. The mutagenized plasmid library is conjugally transferred (i.e., by mating) from an *E. coli* donor into a *P. aeruginosa* recipient strain that expresses the I-*Sce*I enzyme, which prevents stable cointegrate formation. Transposon insertion mutants can be pooled and analyzed by genetic footprinting. Growth of *P. aeruginosa* mutants that have sustained an insertion in an essential gene will be selected against under the specific culture conditions. (Adapted from Ref. ***30***. Copyright © 2004, with permission from Begell House Inc.)

I-*Sce*I sites. Release of the cloned insert provides a linear substrate for integration by homologous recombination into the *P. aeruginosa* chromosome. We observed that expression of I-*Sce*I in *P. aeruginosa* was extremely effective in promoting gene replacement events and strongly selected against cointegrate formation during allelic exchange.

The power of SCE jumping lies in its ability to generate exconjugants that exclusively contain replacements of the endogenous locus with the transposon mutagenized DNA, rather than integration of the delivery plasmid by a single crossover to generate a cointegrate strain, which would retain the wild-type locus. This feature is critical for analysis of essential genes by genetic footprinting. To develop SCE jumping in *P. aeruginosa*, the *pyrF* locus, which encodes an orotidine-5′-phosphate decarboxylase required for biosynthesis of uracil *(31)*, was chosen as a test gene for targeted knockout. The *pyrF* knockout delivery construct containing a replacement of the *pyrF* gene with a gentamicin resistance cassette inserted between *pyrF* flanking regions, all cloned between I-*Sce*I sites, is introduced into *P. aeruginosa* by conjugation. Because *pyrF* mutants require uracil for growth, the desired allelic replacement mutants will grow only on minimal medium containing uracil, whereas cointegrants containing both the mutant DNA and wild-type locus will grow on minimal medium with or without uracil. Therefore, the efficiency of the I-*Sce*I nuclease in promoting gene replacement events with this delivery construct can be assessed by determining the frequency of double-crossover formation versus cointegrate formation via selection on minimal medium in the presence or absence of uracil.

Our results showed that after selection for gentamicin-resistant exconjugants in the presence of uracil, targeted knockout of *pyrF* in a *P. aeruginosa* strain expressing I-*Sce*I resulted in gene replacement at 100% frequency among representative isolates analyzed (28/28). Gentamicin selection in parallel of the same conjugation mixture in the absence of uracil yielded at least a 10,000-fold decrease in the frequency of colony formation *(29)*. This result supports the findings in *E. coli* in which 100- to 1000-fold enhancement of homologous recombination was observed in the resolution of a cointegrate structure mediated by I-*Sce*I–induced double-strand breaks in the *E. coli* chromosome *(32)*. SCE jumping is illustrated in **Figure 3**, and the general steps of this method are as follows:

1. PCR amplify chromosomal region of interest (**Section 3.1.1**).
2. Clone PCR product between I-*Sce*I sites in a delivery vector that does not replicate in *P. aeruginosa*. For example, replication of plasmids containing a ColE1 origin of replication derived from the ColE1 plasmid of *E. coli* *(33, 34)* is not supported in many nonenteric, Gram-negative bacteria, including *P. aeruginosa.*
3. Perform *in vitro* transposition reaction with *mariner* transposon and *Himar1* transposase followed by DNA purification as described in **Section 3.1.1** and **Section 3.1.2**. Single-stranded gaps on either side of the transposon insertion do not need to be repaired with T4 DNA polymerase *in vitro* prior to introduction into *E. coli* in **step 4** below because the gaps are repaired *in vivo* by this bacterium after electroporation.
4. Electroporate *in vitro* mutagenized plasmid pool into an *E. coli* donor strain (e.g., S17-1).
5. Plate mixture onto LB agar containing the appropriate antibiotic marker. This protocol typically generates libraries representing 10^3 to 10^4 different transposon insertion events in the plasmid.

6. Grow to log phase ~10^8 *E. coli* donors carrying a library of the *in vitro* mutagenized plasmid (representing ~1000-fold excess of the 10^3 to 10^4 different transposon insertion events) and ~10^8 *P. aeruginosa* recipients expressing the I-*Sce*I endonuclease.
7. Mix 1 to 5 mL each of log phase donor and recipient cells.
8. Collect conjugation mixture by vacuum filtration onto membrane filters (0.2-µm analytical test filter funnel).
9. Wash cells by aspiration of 10 to 15 mL of 10 mM $MgSO_4$ across the filter.
10. Remove filter and place on LB agar plates, with the side coated with cells facing up.
11. Incubate plates with filters containing cells at 37°C to allow mating to occur (5 h to overnight).
12. After mating, transfer filter to a sterile tube containing LB broth (~1 mL) and vortex to remove mating mixture.
13. Plate out several dilutions of the mating mixture on LB agar containing the appropriate antibiotics.
14. Pool transposon insertion mutants in LB and dilute cells in distilled water to OD_{600} of ~0.1. Use 1 µL of the dilution as template for PCR amplification for genetic footprinting. Alternatively, isolate genomic DNA from transposon mutant pool for genetic footprinting analyses.

3.3. Functional Analysis of Essential Genes in H. influenzae

Understanding the function of essential genes is inherently challenging because, by definition, they are required for optimal growth and viability; therefore, simple knockout experiments are not feasible. However, conditional expression or conditionally active alleles can be used to examine cells during depletion of the essential gene product. Toward this end, we developed two approaches that facilitate functional analyses of essential genes: (1) an inducible expression system utilizing the D-xylose catabolic operon of *H. influenzae* and (2) a marker-linker PCR-mediated mutagenesis method for generation of temperature-sensitive mutations (*35*).

3.3.1. Expression of Essential Genes Using the D-Xylose–Inducible Promoter

A caveat to knocking out many essential genes is that bypass suppressor mutations may occur to allow for growth of a given mutant. The key advantage of the strategy outlined here is that the essential gene of interest is disrupted or deleted in the presence of a conditionally expressed complementing copy of the gene. This conditional expression system utilizes a D-xylose–regulated promoter of the *H. influenzae* D-xylose catabolic operon. Introduction of an essential gene into delivery vector, pXT10, allows sufficient regulated expression of an essential gene under the control of the D-xylose–inducible *xylA* promoter (*35*). The essential gene at its endogenous location can then be disrupted or deleted by standard methods in the presence of D-xylose. Functionality of the essential gene can then be evaluated when the complementing copy of the gene is made conditionally inactive by removal of the inducer. In contrast, presence of xylose induces expression of the essential gene from the *xylA* promoter and allows growth and recovery of the mutant lacking the native copy of the gene. In the expected case, inability of the resulting strain to grow in the absence of xylose verifies the role of the gene in growth or survival. It is possible that pinpoint colonies may be obtained in the absence of inducer. This may be due to low basal induction from the *xylA* promoter as a response to the starvation condition that occurs during the transformation process *(35)*. Alternatively, transformants containing bypass suppressor mutations could arise in the absence of inducer. Evaluating the frequencies of transformants obtained between the plus/minus xylose plates is useful for interpreting whether this might be occurring. The overall scheme is illustrated in **Figure 4** with the following steps:

1. PCR amplify open reading frame (include initiation and termination codon) of essential gene of interest from genomic DNA template with primers containing *Sap*I sites.
2. Purify PCR product using a QIAquick spin column (Qiagen).
3. Digest purified PCR product with *Sap*I.
4. Clone PCR product into the *Sap*I sites of pXT10.
5. Transform resultant construct into *H. influenzae* by natural transformation *(25)* and select for transformants on sBHI agar containing tetracycline overnight at 35°C.
6. Disrupt or delete native copy of the essential gene of interest. This can be achieved by transforming *H. influenzae* with a deletion construct or a PCR product containing an antibiotic marker flanked by homologous sequences corresponding with the locus targeted for deletion.
7. Select for transformants in the presence and absence of 1 mM D(+)-xylose inducer and the appropriate antibiotic in sBHI agar overnight at 35°C.

3.3.2. Isolation of Conditionally Lethal Mutations by Marker-Linked Mutagenesis

This approach utilizes a genomic-scale mutant bank such as the *H. influenzae* mutant library generated by *in vitro* transposon mutagenesis in **Section 3.1**. The overall scheme is illustrated in **Figure 5**, and the steps are outlined below. The general concept of this approach takes advantage of the antibiotic resistance marker provided by *mariner* transposon insertions proximal to an essential gene of interest. Amplification of the region containing the transposon insertion and the essential gene under mutagenic PCR conditions generates random mutations within the PCR product. By varying PCR conditions, the level of mutagenesis can be varied. These mutations can then be introduced into *H. influenzae* using the linked antibiotic resistance marker for selection and the resulting mutants screened to identify those containing conditional lethal mutations in the essential gene. Mutation frequencies within the antibiotic resistance gene itself can be used to gauge the frequency of mutagenesis within the region of interest.

1. Plate dilutions of mutant pool generated in **Section 3.1** corresponding with the chromosomal region containing the gene to be mutagenized to obtain single colonies.
2. Set up reactions in 96-well format for PCR using primers and reaction conditions as described in **Section 3.1.6** with the exception that single colonies are used as template.
3. Using sterile pipette tips, replica-patch 95 single colonies to gridded positions on culture plates, and after patching, touch each tip to the PCR mixture in a different well of the 96-well PCR reaction tube (well number 96 will serve as a “no template control”).
4. Conduct PCR as described for the genetic footprinting method described in **Section 3.1.6**.
5. Analyze PCR reactions by agarose gel electrophoresis.
6. Use genomic map information in conjunction with the PCR product lengths to choose a particular mutant from the analyzed pool that contains a transposon insertion near the essential gene of interest (**Fig. 5**).

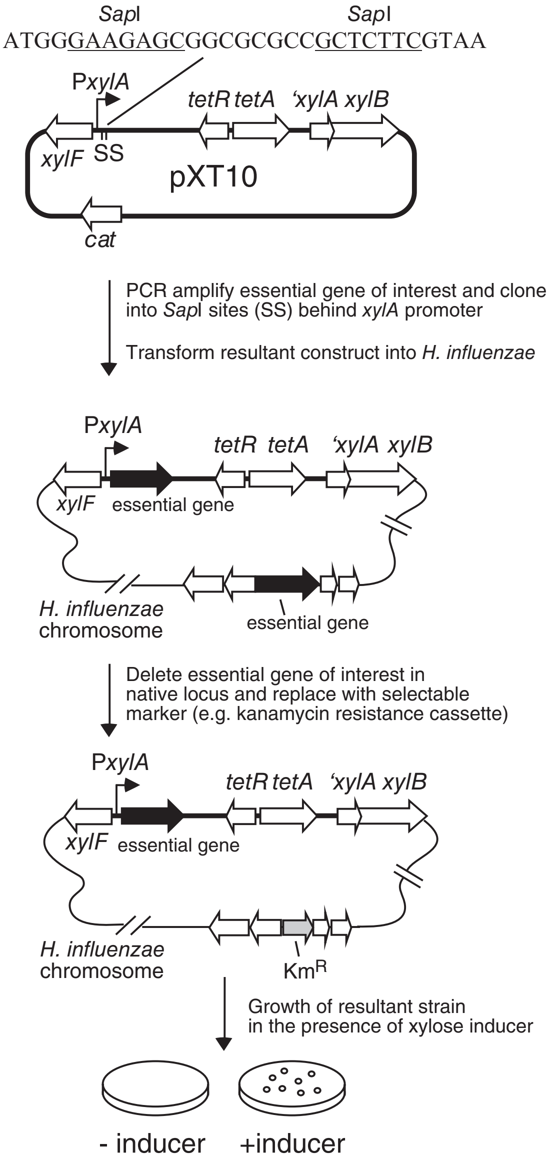

SapI
SapI
ATGGGAAGAGCGGCGCGCCGCTCTTCGTAA
PxylA
tetR tetA
'xylA xylB
xylF
SS
pXT10
cat
PCR amplify essential gene of interest and clone into SapI sites (SS) behind xylA promoter
Transform resultant construct into H. influenzae
PxylA
tetR tetA
'xylA xylB
xylF
essential gene
H. influenzae chromosome
essential gene
Delete essential gene of interest in native locus and replace with selectable marker (e.g. kanamycin resistance cassette)
PxylA
tetR tetA
'xylA xylB
xylF
essential gene
H. influenzae chromosome
Km^R
Growth of resultant strain in the presence of xylose inducer
- inducer
+inducer

Fig. 4. Conditional expression system in *H. influenzae.* The open reading frame of an essential gene of interest is cloned into the *Sap*I sites (SS) of the suicide delivery vector pXT10 immediately downstream of the *xyl*A promoter (P*xylA*). *tetR tetA*, tetracycline resistance locus; *cat*, chloramphenicol resistance gene. The resultant plasmid is transformed into *H. influenzae* for integration into the *xyl* locus. The native copy of the essential gene is targeted for deletion and replaced with an antibiotic marker, for example, kanamycin resistance cassette (Km). The transformation mixture is diluted on sBHI agar containing kanamycin with and without xylose and the frequencies of the number of transformants evaluated between the plus/minus xylose plates. (Adapted from Ref. 35. Copyright © 2003, with permission from Elsevier.)

7. Using this specific transposon insertion mutant as source of template, conduct mutagenic PCR with chromosomal primers that flank the transposon insertion and essential gene of interest. Random mutations within products are generated during PCR in the presence of $MnCl_2$ at a final concentration of 0.0125 mM to 0.125 mM.
8. Transform the mutagenized PCR products into *H. influenzae* and select on sBHI in the presence of Km at 30°C.

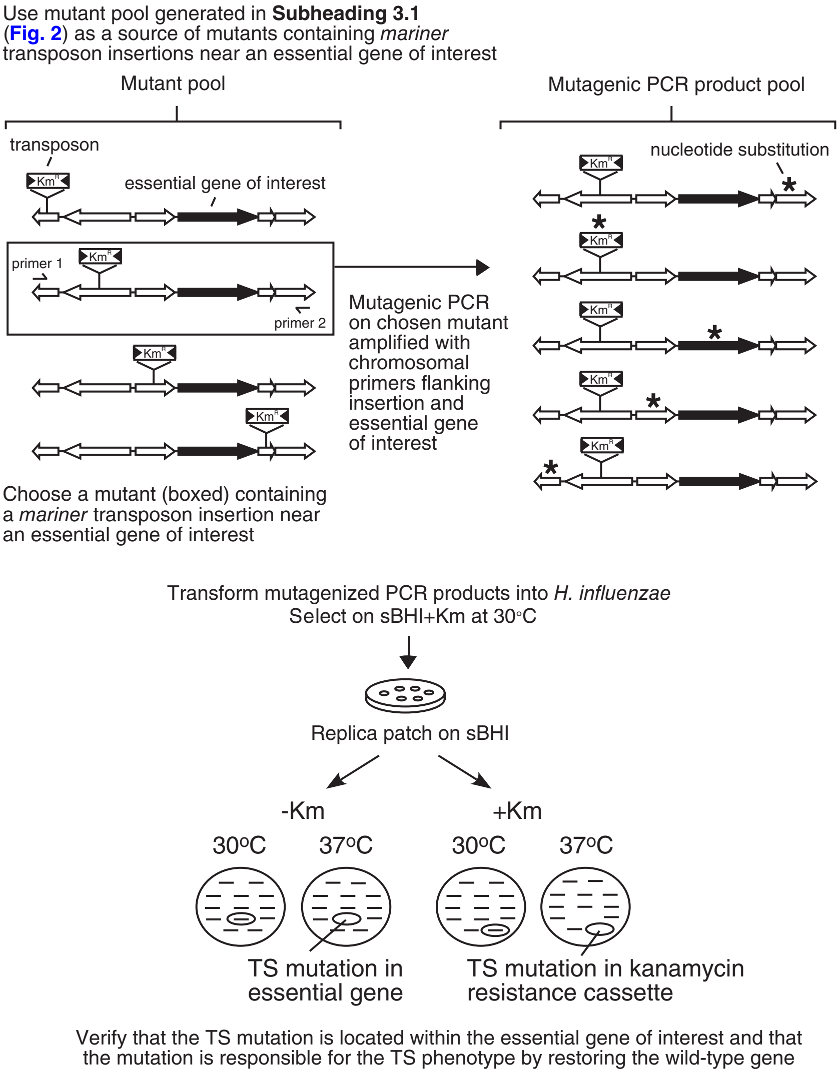

Use mutant pool generated in **Subheading 3.1** (Fig. 2) as a source of mutants containing *mariner* transposon insertions near an essential gene of interest
Mutant pool
Mutagenic PCR product pool
transposon
Km^R
essential gene of interest
nucleotide substitution
primer 1
primer 2
Mutagenic PCR on chosen mutant amplified with chromosomal primers flanking insertion and essential gene of interest
Choose a mutant (boxed) containing a *mariner* transposon insertion near an essential gene of interest
Transform mutagenized PCR products into *H. influenzae*
Select on sBHI+Km at 30°C
Replica patch on sBHI
-Km
+Km
30°C
37°C
30°C
37°C
TS mutation in essential gene
TS mutation in kanamycin resistance cassette
Verify that the TS mutation is located within the essential gene of interest and that the mutation is responsible for the TS phenotype by restoring the wild-type gene

Fig. 5. Mutagenic PCR in *H. influenzae*. Use mutant pool generated in **Section 3.1** as a source of mutants containing *mariner* transposon insertions near an essential gene of interest. Select a particular mutant that contains a transposon insertion proximal to the essential gene and perform mutagenic PCR followed by transformation into *H. influenzae*. Temperature-sensitive (TS) mutants are selected at 30°C. The kanamycin-resistance gene cassette is used to assess frequency of TS mutations generated by mutagenic PCR. (Adapted from Ref. 35. Copyright © 2003, with permission from Elsevier.)

9. Replica-patch Km-resistant transformants onto sBHI plates in the presence and absence of Km at 30°C versus 37°C. Transformants growing at 30°C but not at 37°C in the absence of Km correspond with mutants containing temperature-sensitive mutations in the essential gene of interest.
10. Transformants growing at 30°C but not at 37°C in the presence of Km correspond with mutants containing TS mutations in the Km gene cassette. Compare the number of transformants that grow in the presence of Km at 30°C versus 37°C to evaluate the frequency of the TS mutations generated by mutagenic PCR in the Km gene. This frequency can be used to standardize the level of mutagenesis between experiments.
11. For verification that the TS mutation is located within the essential gene of interest and that the mutation is responsible for the TS phenotype, prepare a naturally competent culture of the TS mutant at 30°C for transformation with the wild-type gene cloned in the pXT10 delivery vector, linear DNA fragments corresponding with the essential gene, or a replicating plasmid, for example, pGJB103 ***(25)*** containing a cloned copy of the essential gene.
12. Incubate the competent cells with the complementing construct, wild-type DNA fragment, or without added DNA at 30°C for 1 h.
13. Shift cultures to 37°C with shaking and monitor 6 to 12 h for restoration of growth in comparison with the control culture that did not receive the wild-type gene ***(35)***.

Notes

1. Design principles for complementation of essential genes.

   Once an inducible expression system is generated for an essential gene (**Section 3.3.1**), then it is possible to modify this system for constitutive complementation that does not require an inducer. Constitutive complementation with the essential gene's endogenous promoter, for example, may be desirable for cases in which the D-xylose–inducible copy of the essential gene does not fully restore the parental phenotype. The inducible copy can be replaced with a constitutively expressed copy of the essential gene under the transcriptional control of its own promoter. Briefly, a PCR fragment of the essential gene of interest containing its native promoter is cloned into the *Sap*I sites of a pXT10 derivative containing an antibiotic marker other than Tet^R followed by transformation into *H. influenzae* to replace the conditionally expressed copy. Note that transformation is performed in the presence of the conditional active essential protein such that the constitutively complementing copy is transformed into a recipient strain verified to display the mutant phenotype in the absence of conditional complementation. In this way, the mutant phenotype can be observed in the absence of D-xylose in the strain containing the D-xylose–inducible essential gene, and complementation with the essential gene driven by its native promoter should restore the parental phenotype, ruling out second site mutations or transcriptional polarity of the knockout mutation as causes of the phenotype.
2. Functional analysis of minimal *Himar1 mariner* inverted repeat sequences. *Mariner*-based transposons have proved highly efficient at both *in vitro* and *in vivo* transposon mutagenesis of bacteria as first shown in the two naturally competent bacteria, *H. influenzae* and *Streptococcus pneumoniae* ***(20)***, and then in two bacteria that are not naturally transformable, *Mycobacterium smegmatis* and *E. coli* ***(36)***, respectively. Subsequently, *mariner*-based transposons have been engineered to function as effective genetic tools in a growing list of bacterial systems including *Campylobacter jejuni* ***(27)***, *Vibrio cholerae* ***(16)***, and *P. aeruginosa* ***(29)***.

   Use of the *mariner* transposon would be ideal for creating reporter gene fusions in chromosomal regions of interest. To facilitate the creation of useful *mariner* transposon derivatives, we sought to determine the minimum *Himar1* inverted terminal repeat sequences required for efficient transposition (**Fig. 6**). We engineered a *mariner* transposon derivative that includes a cloned copy of the *Himar1* transposase with the *Himar1* repeats flanking *aacC1*, a gentamicin resistance cassette. The parent construct (pBC KS+ derivative) contains the *Himar1* 31-bp imperfect inverted repeat (a single G/A transition at position 28 of the terminal inverted repeat), with the first 27 bp perfectly inverted (**Fig. 6**). Plasmid derivatives containing deletions and/or sequence substitutions within the terminal inverted repeat were engineered and then tested in a "mating out" assay in *E. coli* to assess transposition efficiencies. In the mating-out assay, the *mariner* arm deletion plasmids carrying the *Himar1* transposase gene are each introduced into the *E. coli* SM10 donor (or mobilizing) strain that contains the transfer genes of the broad host range IncP type plasmid, RP4, integrated in its chromosome ***(37)***. This donor strain also carries a "target" plasmid carrying an ampicillin resistance cassette and an origin of transfer, allowing it to be mobilized from SM10 to a recipient strain. The efficiency of *in vivo mariner* transposition is quantified via conjugal transfer of the target plasmid from SM10 into TOP10 *E. coli* (Invitrogen, Carlsbad, CA) and selection for the transposon-encoded resistance determinant. **Figure 6** shows the efficiency of transposition relative to the *Himar1* arm repeat length. The results indicate that the first

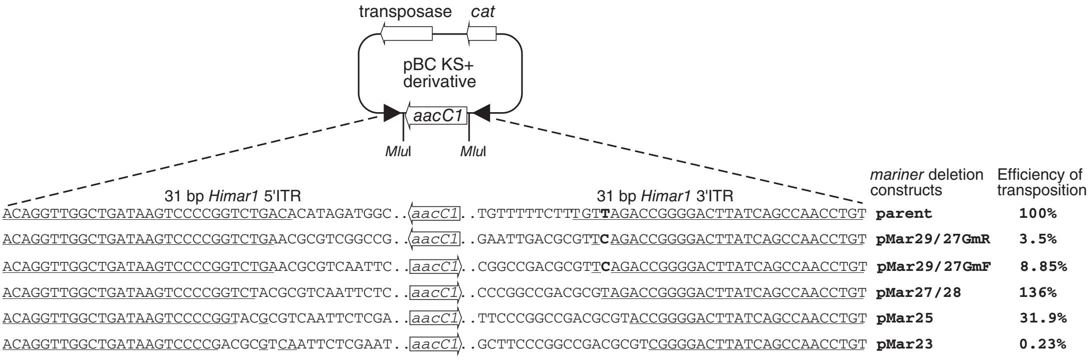

transposase
cat
pBC KS+ derivative
aacC1
MluI
MluI
31 bp Himar1 5'ITR
31 bp Himar1 3'ITR
mariner deletion constructs
Efficiency of transposition
ACAGGTTGGCTGATAAGTCCCCGGTCTGACACATAGATGGC..aacC1..TGTTTTTCTTTGTTAGACCGGGGACTTATCAGCCAACCTGT parent 100%
ACAGGTTGGCTGATAAGTCCCCGGTCTGAACGCGTCGGCCG..aacC1..GAATTGACGCGTTCAGACCGGGGACTTATCAGCCAACCTGT pMar29/27GmR 3.5%
ACAGGTTGGCTGATAAGTCCCCGGTCTGAACGCGTCAATTC..aacC1..CGGCCGACGCGTTCAGACCGGGGACTTATCAGCCAACCTGT pMar29/27GmF 8.85%
ACAGGTTGGCTGATAAGTCCCCGGTCTACGCGTCAATTCTC..aacC1..CCCGGCCGACGCGTAGACCGGGGACTTATCAGCCAACCTGT pMar27/28 136%
ACAGGTTGGCTGATAAGTCCCCGGTACGCGTCAATTCTCGA..aacC1..TTCCCGGCCGACGCGTACCGGGGACTTATCAGCCAACCTGT pMar25 31.9%
ACAGGTTGGCTGATAAGTCCCCGACGCGTCAATTCTCGAAT..aacC1..GCTTCCCGGCCGACGCGTCGGGGACTTATCAGCCAACCTGT pMar23 0.23%

Fig. 6. Functional analysis of minimal *Himar1 mariner* repeat sequences. A *mariner* transposon delivery plasmid contains a copy of the *Himar1* transposase for *in vivo* transposition, a chloramphenicol (*cat*) resistance gene, and a gentamicin resistance cassette (*aacC1*) cloned between *Mlu*I restriction sites flanked by sequences containing the 31-bp imperfect *Himar1* inverted terminal repeats (ITR) (underlined). *Himar1* ITR nucleotide sequences that remained unchanged in the *mariner* arm deletion constructs are underlined. pMar29/27 GmR and GmF contain a T to C nucleotide (in boldface) change at position 28 in the *Himar1* 3′ ITR. Transcription of *aacC1* in pMar29/27 GmR and pMar29/27 GmF is in the opposite orientation (represented by the arrow). Transposition efficiency corresponding with each deletion construct is compared with that of the parent construct. Percentage of efficiency is derived from the number of transconjugants in a "mating out" assay (described in text) normalized to the number of donor cells divided by the number of recipient cells. Number of transconjugants (CFU/mL) from two independent mating-out assays was recorded for parent (1.5 × 10^4 and 0.83 × 10^4), pMar27/28 (2.1 × 10^4 and 1 × 10^4), and pMar23 (2.3 × 10^1 and 3.1 × 10^1) constructs. The number of transconjugants (CFU/mL) from a single mating assay was performed with pMar29/27 GmR (4 × 10^2), pMar29/27 GmF (1 × 10^3), and pMar25 (3.7 × 10^3). In cases where two independent mating-out assays were performed for a deletion construct, the average of the normalized numbers of transconjugants was used to calculate the transposition efficiency.

27 bp and 28 bp of the *Himar1* 5′ITR and 3′ITR, respectively, are sufficient for efficient transposition. However, transposition with the first 25 bp of the repeat decreases to ~32% of the wild-type level, with transposition virtually abolished with only the first 23 bp of the repeat present. Of note, transposon constructs, pMar29/27 GmR and pMar29/27 GmF, which contain the first 29 bp of the *Himar1* repeat but with a single nucleotide change from T to C at position 28 at the 3′ITR, show drastically reduced transposition efficiency. This nucleotide change may have affected the binding efficiency of the *Himar1* transposase (*38, 39*).

Acknowledgments

We thank Dr. Christopher Sassetti and Jane Harrington for helpful comments on the manuscript. This work was supported in part by grants from the NIH (AI49437) and the American Heart Association to B.J.A.; in part, a grant from the Cystic Fibrosis Foundation to S.M.S.W.

References

1. Kang, Y., Durfee, T., Glasner, J. D., Qiu, Y., Frisch, D., Winterberg, K. M., et al. (2004) Systematic mutagenesis of the *Escherichia coli* genome. *J. Bacteriol.* **186**, 4921–4930.
2. Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M. R., Maruf, M., et al. (2006) Essential genes of a minimal bacterium. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 425–430.
3. Liberati, N. T., Urbach, J. M., Miyata, S., Lee, D. G., Drenkard, E., Wu, G., et al. (2006) An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2833–2838.
4. Suzuki, N., Okai, N., Nonaka, H., Tsuge, Y., Inui, M., and Yukawa, H. (2006) High-throughput transposon mutagenesis of *Corynebacterium glutamicum* and construction of a single-gene disruptant mutant library. *Appl. Environ. Microbiol.* **72**, 3750–3755.
5. Holtman, C. K., Chen, Y., Sandoval, P., Gonzales, A., Nalty, M. S., Thomas, T. L., et al. (2005) High-throughput functional analysis of the *Synechococcus elongatus* PCC 7942 genome. *DNA Res.* **12**, 103–115.
6. Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2001) Comprehensive identification of conditionally essential genes in mycobacteria. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 12712–12717.
7. Tong, X., Campbell, J. W., Balazsi, G., Kay, K. A., Wanner, B. L., Gerdes, S. Y., et al. (2004) Genome-scale identification of conditionally essential genes in *E. coli* by DNA microarrays. *Biochem. Biophys. Res. Commun.* **322**, 347–354.
8. Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balazsi, G., Ravasz, E., Daugherty, M. D., et al. (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684.
9. Jacobs, M. A., Alwood, A., Thaipisuttikul, I., Spencer, D., Haugen, E., Ernst, S., et al. (2003) Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 14339–14344.
10. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., et al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, E1–E11.
11. Ji, Y., Woodnutt, G., Rosenberg, M., and Burnham, M. K. (2002) Identification of essential genes in *Staphylococcus aureus* using inducible antisense RNA. *Methods Enzymol.* **358**, 123–128.

12. Forsyth, R. A., Haselbeck, R. J., Ohlsen, K. L., Yamamoto, R. T., Xu, H., Trawick, J. D., et al. (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol. Microbiol.* **43**, 1387–1400.
13. Lehoux, D. E., Sanschagrin, F., and Levesque, R. C. (2002) Identification of in vivo essential genes from *Pseudomonas aeruginosa* by PCR-based signature-tagged mutagenesis. *FEMS Microbiol. Lett.* **210**, 73–80.
14. Chalker, A. F., Minehart, H. W., Hughes, N. J., Koretke, K. K., Lonetto, M. A., Brinkman, K. K., et al. (2001) Systematic identification of selective essential genes in *Helicobacter pylori* by genome prioritization and allelic replacement mutagenesis. *J. Bacteriol.* **183**, 1259–1268.
15. Song, J. H., Ko, K. S., Lee, J. Y., Baek, J. Y., Oh, W. S., Yoon, H. S., et al. (2005) Identification of essential genes in *Streptococcus pneumoniae* by allelic replacement mutagenesis. *Mol. Cells* **19**, 365–374.
16. Judson, N., and Mekalanos, J. J. (2000) TnAraOut, a transposon-based approach to identify and characterize essential bacterial genes. *Nat. Biotechnol.* **18**, 740–745.
17. Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., Arnaud, M., et al. (2003) Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4678–4683.
18. Salama, N. R., Shepherd, B., and Falkow, S. (2004) Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori*. *J. Bacteriol.* **186**, 7926–7935.
19. Smith, V., Botstein, D., and Brown, P. O. (1995) Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 6479–6483.
20. Akerley, B. J., Rubin, E. J., Camilli, A., Lampe, D. J., Robertson, H. M., and Mekalanos, J. J. (1998) Systematic identification of essential genes by *in vitro mariner* mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 8927–8932.
21. Akerley, B. J., Rubin, E. J., Novick, V. L., Amaya, K., Judson, N., and Mekalanos, J. J. (2002) A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 966–971.
22. Akerley, B. J., and Lampe, D. J. (2002) Analysis of gene function in bacterial pathogens by GAMBIT. *Methods Enzymol.* **358**, 100–108.
23. Smith, H. O., Tomb, J. F., Dougherty, B. A., Fleischmann, R. D., and Venter, J. C. (1995) Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. *Science* **269**, 538–540.
24. Hosking, S. L., Deadman, M. E., Moxon, E. R., Peden, J. F., Saunders, N. J., and High, N. J. (1998) An *in silico* evaluation of Tn*916* as a tool for generalized mutagenesis in *Haemophilus influenzae* Rd. *Microbiology* **144** (Pt 9), 2525–2530.
25. Barcak, G. J., Chandler, M. S., Redfield, R. J., and Tomb, J. F. (1991) Genetic systems in *Haemophilus influenzae*. *Methods Enzymol.* **204**, 321–342.
26. Smith, V., Chou, K. N., Lashkari, D., Botstein, D., and Brown, P. O. (1996) Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* **274**, 2069–2074.
27. Hendrixson, D., Akerley, B., and DiRita, V. (2001) Transposon mutagenesis of *Campylobacter jejuni* identifies a bipartite energy taxis system required for motility. *Mol. Microbiol.* **40**, 214–224.
28. Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., et al. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406**, 959–964.

29. Wong, S. M., and Mekalanos, J. J. (2000) Genetic footprinting with *mariner*-based transposition in *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 10191–10196.
30. Wong, S. M. (2004) SCE jumping: genetic tool for allelic exchange in bacteria. *Crit. Rev. Eukaryot. Gene Expr.* **14**, 53–64.
31. Strych, U., Wohlfarth, S., and Winkler, U. K. (1994) Orotidine-5′-monophosphate decarboxylase from *Pseudomonas aeruginosa* PAO1: cloning, overexpression, and enzyme characterization. *Curr. Microbiol.* **29**, 353–359.
32. Pósfai, G., Kolisnychenko, V., Bereczki, Z., and Blattner, F. R. (1999) Markerless gene replacement in *Escherichia coli* stimulated by a double-strand break in the chromosome. *Nucleic Acids Res.* **27**, 4409–4415.
33. Bazaral, M., and Helinski, D. R. (1968) Circular DNA forms of colicinogenic factors E1, E2 and E3 from *Escherichia coli*. *J. Mol. Biol.* **36**, 185–194.
34. Konisky, J. (1982) Colicins and other bacteriocins with established modes of action. *Annu. Rev. Microbiol.* **36**, 125–144.
35. Wong, S. M., and Akerley, B. J. (2003) Inducible expression system and marker-linked mutagenesis approach for functional genomics of *Haemophilus influenzae*. *Gene* **316**, 177–186.
36. Rubin, E. J., Akerley, B. J., Novik, V. N., Lampe, D. J., Husson, R. N., and Mekalanos, J. J. (1999) *In vivo* transposition of *mariner*-based elements in enteric bacteria and mycobacteria. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 1645–1650.
37. Simon, R., Priefer, U., and Puhler, A. (1983) A broad host range mobilization system for *in vivo* genetic engineering: transposon mutagenesis in gram negative bacteria. *Biotechnology* **1**, 784–791.
38. Lampe, D. J., Churchill, M. E., and Robertson, H. M. (1996) A purified *mariner* transposase is sufficient to mediate transposition *in vitro*. *EMBO J.* **15**, 5470–5479.
39. Lipkow, K., Buisine, N., Lampe, D. J., and Chalmers, R. (2004) Early intermediates of *mariner* transposition: catalysis without synapsis of the transposon ends suggests a novel architecture of the synaptic complex. *Mol. Cell Biol.* **24**, 8301–8311.

4

Transposon Site Hybridization in *Mycobacterium tuberculosis*

Jeffrey P. Murry, Christopher M. Sassetti, James M. Lane, Zhifang Xie, and Eric J. Rubin

Summary

Microarray mapping of transposon insertions can be used to quantify the relative abundance of different transposon mutants within a complex pool after exposure to selective pressure. The transposon site hybridization (TraSH) method applies this strategy to the study of *Mycobacterium tuberculosis* and can be adapted to the study of other microorganisms. This chapter describes the methods used to mutagenize mycobacteria with transposons, extract genomic DNA, amplify genomic DNA adjacent to transposon ends using polymerase chain reaction and T7 transcription, and synthesize labeled cDNA. It also describes methods used to construct an appropriate microarray, hybridize labeled cDNA, and analyze the microarray data. Important considerations involved in the experimental design of the selective pressure, the design of the microarray, and the statistical analysis of collected data are discussed.

Key Words: method design; microarray; *Mycobacterium tuberculosis*; transposon; TraSH.

1. Introduction

The use of random mutagenesis in combination with microarray technology has enabled the development of methods that allow comprehensive identification of genetic elements required for bacterial replication under various selective conditions (*1, 2*). Transposon site hybridization (TraSH) was developed using these technologies to quantify relative abundance of transposon mutants in the context of a complex pool. In this method, genomic DNA from the transposon pool is digested and ligated to an adaptor. Genomic regions adjacent to the transposon insertions are specifically amplified using polymerase chain reaction (PCR) and an outward-facing T7 promoter on the transposon. Labeling efficiency is increased by synthesizing cDNA from the transcription products, and the resulting products are quantified by hybridization to microarrays. As TraSH was specifically developed to study *Mycobacterium tuberculosis*, this organism is used to illustrate the method. Similar methods have been developed for other organisms (*1, 3, 4*). These methods have been successfully used to identify genetic elements important for growth *in vitro* and in mouse and macrophage models of infection (*1, 3–8*).

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

2. Materials

1. ϕMycoMarT7 transposon donor phasmid.
2. *Mycobacterium smegmatis* strain mc^{2}155 and *Mycobacterium tuberculosis* H37Rv or *Mycobacterium bovis* BCG.
3. Middlebrook 7H9 broth (BD Difco, Franklin Lakes, NJ): unless otherwise noted, 1 L 7H9 contains 2 mL glycerol, 0.05% (v/v) Tween-80, and 100 mL Middlebrook OADC in water.
4. Middlebrook OADC (BD BBL) contains the following components: 8.5 g NaCl, 50 g bovine albumin (fraction V), 20 g dextrose, 0.03 g catalase, 0.6 mL oleic acid per liter of water).
5. Middlebrook 7H10 agar (BD Difco): 1 L 7H10 contains 5 mL glycerol and 100 mL OADC in water.
6. 0.2-µm syringe filter and syringe.
7. MP buffer: 50 mM Tris-HCl, pH 7.5 at room temperature, 150 mM NaCl, 10 mM $MgSO_4$, 2 mM $CaCl_2$.
8. Top agar: 0.6% agar (w/v) in 2 mM $CaCl_2$ (add $CaCl_2$ after autoclaving).
9. Kanamycin.
10. 4-mm glass beads (Stern, Walter, Inc., Port Washington, NY).
11. TE buffer: 10 mM Tris-HCl, pH 8.0 and 1 mM EDTA; sterilize by autoclaving.
12. Chloroform and methanol (**Note 1**).
13. Lysozyme: 10 mg/mL stock (Sigma-Aldrich, St. Louis, MO).
14. Proteinase K: 10 mg/mL stock (New England Biolabs, Beverly, MA).
15. Sodium dodecyl sulfate (SDS).
16. Phenol (with isoamyl alcohol; **Note 1**).
17. Isopropanol and sodium acetate.
18. 70% ethanol.
19. Agarose and gel electrophoresis equipment.
20. QIAquick gel extraction, QIAquick PCR purification, and RNeasy mini kits (Qiagen, Valencia, CA).
21. Taq DNA polymerase, restriction enzymes, T4 DNA ligase (New England Biolabs).
22. Oligonucleotide primers (Integrated DNA Technologies, Coralville, IA).
23. Dimethyl sulfoxide (DMSO).
24. $MgCl_2$.
25. Deoxynucleotide solution mix (dATP, dCTP, dTTP, and dGTP).
26. AmpliTaq Gold DNA polymerase (Applied Biosystems, Foster City, CA) and Pfu DNA polymerase (Stratagene, La Jolla, CA).
27. SYBR Green I nucleic acid gel stain (light sensitive; Molecular Probes, Eugene, OR).
28. Real-time thermal cycling system.
29. Vacufuge.
30. MEGAShortScript high-yield transcription kit (Ambion, Austin, TX).
31. RNaseOUT recombinant ribonuclease inhibitor (Invitrogen, Carlsbad, CA).
32. 10× aa-dNTP mix (store at –80°C): 5 mM dATP, 5 mM dCTP, 5 mM dGTP, 2 mM dTTP, 3 mM aminoallyl dUTP (Sigma-Aldrich).
33. Superscript III reverse transcriptase (Invitrogen) and 10× first strand buffer: 250 mM Tris-HCl (pH 8.3 at room temperature), 375 mM KCl, 15 mM $MgCl_2$.
34. Ethylenediaminetetraacetic acid (EDTA), NaOH, and HCl.
35. cDNA wash buffer: NaCl 0.58 g, H_2O 20 mL, and ethanol 80 mL.
36. Cy3 and Cy5 monoreactive dyes (light sensitive; GE Healthcare, Giles, UK).

37. Multiscreen PCR purification plates (Millipore, Billerica, MA).
38. CodeLink activated slides (GE Healthcare, Giles, UK).
39. Tecan HS400 hybridization station (Tecan, Grödig, Austria).
40. SSC: 0.15 M sodium chloride/0.015 M sodium citrate, pH 7.
41. Prehybridization buffer: 5× SSC, 0.1% SDS, 0.1% bovine serum albumin, 100 μg/mL yeast tRNA (Invitrogen).
42. Hybridization buffer: 5× SSC, 0.1% SDS, 50% formamide, 200 μg/mL yeast tRNA.

3. Methods

The following methods are described below: (1) transposon library construction, (2) transposon mutant selection, (3) preparation of chromosomal DNA from mutant pool, (4) preparation of labeled cDNA, (5) microarray construction, and (6) microarray hybridization.

3.1. Transposon Library Construction

The construction of transposon mutant libraries is described in **Section 3.1.1** to **Section 3.1.4**. This includes descriptions of the transposon vector, the transduction process used to introduce the phage into mycobacteria, and the cultivation and maintenance of the library.

3.1.1. ϕMycoMarT7 Transposon Donor Phasmid

The ϕMycoMarT7 phasmid ***(2)*** contains the highly active C9 *Himar1* transposase gene and the MycoMarT7 transposon on the temperature-sensitive phasmid ϕAE87 ***(9)***. The ϕAE87 phasmid, which was developed in Bill Jacobs' laboratory, efficiently produces phage in *M. smegmatis* at 30°C but does not replicate at 37°C ***(9)***. The C9 *Himar1* transposase is a hyperactive mutant of an enzyme that was originally cloned from the horn fly *Haematobia irritans* and is expressed in the ϕMycoMarT7 phasmid from a mycobacterial promoter ***(10, 11)***. The MycoMarT7 transposon encodes a kanamycin resistance gene, the R6K replication origin, two outward-facing T7 promoters, and two flanking 29-bp inverted repeats that are recognized by the *Himar1* transposase. The sequence of the MycoMarT7 transposon has been deposited in GeneBank (accession no. AF411123). The kanamycin resistance gene allows selection in both mycobacteria and *Escherichia coli*. The R6K replication origin is functional in *pir*+ *E. coli* strains, allowing recovery of the transposon after insertion into target strains. The T7 promoters are oriented so that they drive transcription into adjacent chromosomal DNA. These features make this phasmid suitable for TraSH.

3.1.2. Mycobacterial Phage Stock Preparation

The phage stock used to make the transposon library should be generated from a single temperature-sensitive clone. The following steps can be used to generate a suitable phage stock.

1. After acquiring ϕMycoMarT7, make 10-fold dilutions of the given aliquot in 50 μL MP buffer. Add each dilution to 100 μL of *M. smegmatis* that has been washed twice with 7H9 containing glycerol and ADC (similar to OADC but without oleic acid) but no Tween-80 (**Note 2**). Add the mixture of phage with bacteria to 3.5 mL of top agar (cooled to 42°C)

and pour on a 15-cm LB plate. Incubate at 30°C for about 48 h. A few large plaques may appear earlier, but allow 48 h for the appearance of smaller plaques.

2. Patch several plaques onto two plates containing *M. smegmatis* in top agar as in **step 1**. Incubate one at 30°C and one at 37°C for 1.5 to 2 days. Most if not all isolates should form plaques only at 30°C.
3. Excise agar containing a clone that forms a plaque only at 30°C and crush it in MP buffer. Pellet the agar by centrifugation and titer the supernatant as in **step 1** to determine a dilution that results in nearly confluent plaques.
4. Wash 500 µL of stationary-phase *M. smegmatis* twice with 7H9 containing glycerol and ADC (without Tween). Add enough phage to the cells to create nearly confluent plaques. Add 100 µL of this mixture to 3.5 mL top agar (cooled to 42°C) and pour on a 10-cm 7H10 plate (**Note 3**). Prepare five plates in this manner.
5. Incubate five plates at 30°C until "lacy" (about 2 days), and flood each plate with 3 mL MP buffer. Gently rock plates at 4°C for several hours or overnight, then collect the plate stock and pass over a 0.2-µm syringe filter.

3.1.3. Titering Phage Stock

1. Prepare lawn of *M. smegmatis* by adding 250 µL of stationary-phase culture to 3.5 mL top agar and pouring on an LB plate. Allow this plate to dry for a few hours.
2. Prepare 10-fold dilutions of phage stock in 100 µL MP buffer. Spot 10 µL of each dilution onto the plate and allow the spots to dry. Incubate at 30°C for 2 days and count plaques. The stock titer should be at least 5×10^{10} plaque forming units/mL.

3.1.4. Transduction of M. tuberculosis *or* M. bovis *BCG*

1. Grow *M. tuberculosis* (**Note 4**) or *M. bovis* BCG in a roller bottle with 100 mL 7H9 containing glycerol, OADC, and 0.05% Tween-80 until OD_{600} reaches between 0.8 and 1.0.
2. Spin down 50 mL of culture and wash with 7H9+glycerol and OADC (No Tween, MP buffer can also be used). Resuspend in 5 mL of wash medium and remove an aliquot to serve as a control.
3. Add ~10^{11} phage (or MP buffer to the control) and incubate for 3 to 4 h at 37°C. Freeze the transductants at −80°C in multiple aliquots.
4. Thaw an aliquot of transduced bacilli and plate serial dilutions on 7H10 plates containing 20 µg/mL kanamycin to titer the library.
5. Plate at least 100,000 transductants on several 15-cm 7H10 plates containing 20 µg/mL kanamycin at a density of 20,000 CFU/plate.

3.2. Transposon Mutant Selection

Described below are important principles to be considered when designing selection of transposon mutants and the steps used to recover bacteria after selective pressure.

3.2.1. Design of Selective Pressure

The most important part of a TraSH experiment is the initial selection. Of course, selective pressure was applied to the library when it was originally plated on media as in **step 5** of **Section 3.1.4**. Mutations that produce lethal insertions will not survive this initial outgrowth. The first application of the TraSH method compared a transposon library that was plated on 7H10 media immediately after transduction with that plated on 7H11 media, which contains additional amino acids and supplements ***(2)***. A more difficult comparison was used to identify genes essential for normal *in vitro* replication ***(7)***. In this experiment, labeled genomic DNA was compared with labeled cDNA made from bacilli grown on 7H10. As this experiment did not directly compare a mutagenized library selected under one condition to that in another, it did not have internal controls for transposon insertional bias and required more stringent statistical analysis as discussed below. More recent applications of the TraSH method have compared mutagenized libraries grown under different selective pressures directly with each other, using growth on 7H10 as a control condition ***(5, 6)***.

When designing selective conditions, it is important to consider the magnitude of the expected enrichment. TraSH has been consistently used to detect transposon mutants that are tenfold less abundant in one condition relative to another ***(6)***. Smaller differences may be more difficult to assess with confidence using this method, although more subtle phenotypes can be magnified using serial rounds of selection ***(5, 7)***. In our experience, experimental conditions are more likely to produce detectable enrichment when they allow multiple rounds of replication or significant bacterial death while under selective pressure.

3.2.2. Plating Transposon Libraries

The completion of TraSH methodology requires a fairly large amount of genomic DNA. For this reason, it is helpful to amplify the library by plating it and allowing the formation of individual colonies. During this step, it is important to minimize competition between clones by preventing colonies from overlapping with each other as much as possible. We usually spread at least 10 plates with 20,000 colony-forming units (CFU) of each *M. tuberculosis* transposon library:

1. Add 0.5 mL of library at 40,000 CFU/mL to the surface of a 15-cm plate containing 7H10 media and about 2 dozen sterile 4-mm glass beads (**Note 5**).
2. Shake the plates to evenly spread the bacteria and allow them to dry before removing the glass beads. Incubate the plates at 37°C for 18 to 21 days to allow colony formation.

3.3. Preparation of Chromosomal DNA from Mutant Pool

The following method for the purification of chromosomal DNA is adapted from Belisle and Sonnenberg ***(12)***. We use biosafety level 3 containment for **step 1** to **step 8**, although it is likely that the bacilli are inactivated by **step 2**.

1. Harvest colonies from ten 15-cm plates by scraping into 7H9 medium. Centrifuge the suspension at 3300 × *g* for 10 min at room temperature. Discard the supernatant and resuspend the pellet in 5 mL10 mM Tris-HCl, 1 mM EDTA at pH 9.
2. Mix the resuspended cells with an equal volume of chloroform: methanol (2 : 1) and rock for 5 min.
3. Centrifuge the suspension at 3300 × *g* for 10 min at room temperature. Remove both the aqueous and the organic phases into a 50-mL conical tube.
4. Dry the solid bacterial mass by leaving the tube open in the biosafety cabinet for about 2 h.
5. Add 10 mL TE containing 0.1 M Tris-HCl at pH 9 to the pellet and vortex to resuspend.
6. Add 0.01 volume of 10 mg/mL lysozyme and incubate overnight at 37°C.

7. Add 1 mL 10% SDS. Add proteinase K to a final concentration of 100 µg/mL and vortex the samples. Incubate at 50°C for 3 h.
8. Transfer the viscous solution into a clean tube containing an equal volume of phenol: chloroform (1 : 1). Mix well and let stand for 30 min.
9. Rock the tube for 30 min at room temperature. Centrifuge at 12,000 × *g* for 15 min. Remove the upper aqueous phase to a new tube with an equal volume of chloroform and repeat the centrifugation.
10. Remove the upper aqueous phase to a new tube with an equal volume of isopropanol and 1/10 volume of 3 M sodium acetate (pH 5.2). Spool out the DNA, wash with 70% ethanol, and dissolve in 0.5 to 1 mL TE.

3.4. Preparation of Labeled cDNA

The preparation of labeled cDNA is described in **Section 3.4.1** to **Section 3.4.4**. This includes the partial digestion of chromosomal DNA and adapter ligation, amplification of the regions adjacent to the transposon-insertion sites by PCR, *in vitro* transcription of the PCR products, and synthesis and labeling of cDNA.

3.4.1. Partial Digestion and Adapter Ligation

1. For each sample, mix ~2 µg genomic DNA, enzyme buffer, and water in a total volume of 130 µL. Aliquot the mix into 2 series of six tubes each, putting 15 µL in the first tube and 10 µL in each of the five remaining tubes (**Note 6**).
2. Into each tube with 15 µL, add either 5 U HinP1I or MspI (**Note 7**). Make threefold serial dilutions by taking 5 µL from the first tube with 15 µL and adding it to the next tube, which should have only 10 µL, then mixing the new dilution. Incubate the reaction at 37°C for 1 h, then inactivate the enzymes by incubating at 65°C for 20 min.
3. Run each reaction on a 1% agarose gel (**Fig. 1**). Pick 2 to 4 reactions from each sample that show a homogenous smear from 500 to 2000 bp and cut out this region of the gel

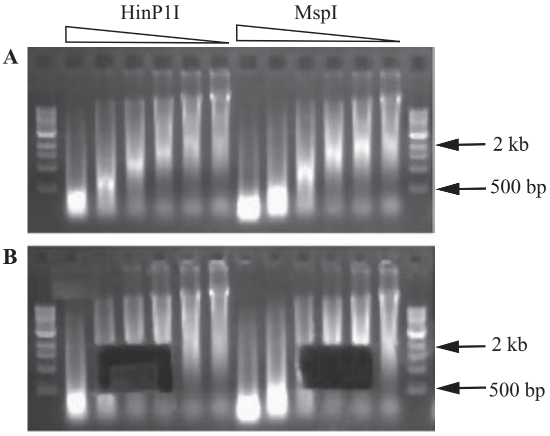


Fig. 1. Partial digestion of genomic DNA. Genomic DNA extracted from a pool of *M. tuberculosis* transposon mutants was digested with serial dilutions of HinP1I or MspI for 1 h at 37°C. **(A)** Digest products were resolved on a 1% agarose gel. **(B)** Fragments between 500 bp and 2 kb in size were extracted and purified.

(**Note 8**). Ideally, each sample would have a similar smearing pattern for the excised reactions, although some variation is inevitable. Purify DNA using the QIAquick gel extraction kit. Elute DNA in 30 µL 2 mM Tris-HCl, pH 8.5.

4. Quantify eluted DNA products using a spectrophotometer or fluorometer and mix equal amounts of DNA from each digestion. Use a vacufuge to dry the mixed product to a total volume of 27 µL.
5. Mix equal volumes of 100-µM solutions of the following adapter oligonucleotides: CGACCACGACCA (includes 3′ C6-TFA-amino modification) and AGTCTCGCA GATGATAAGGTGGTCGTGGT. Heat to 95°C for 5 min, then decrease by 0.1°C/s to 25°C to allow the oligos to anneal to each other.
6. Mix the following for adapter ligation: 27 µL DNA fragments (**step 4**), 4 µL 10× T4 DNA ligase buffer, 8 µL annealed adapter (50 µM each), and 1 µL T4 DNA ligase. Incubate at 16°C overnight.

3.4.2. PCR Amplification of Transposon Ends and Adjacent Chromosomal DNA

PCR is used to amplify the regions adjacent to the transposon-insertion sites. For each sample, two separate PCR reactions containing an adapter-specific primer and a transposon-specific primer are performed. The two transposon primers are specific for different transposon ends so that the transposon junction from each side is amplified separately.

1. For each sample, make two sets of 50-µL PCR reactions containing the following: 1.5 mM $MgCl_2$, 0.25 mM each dNTP, 10% DMSO, 1 µM adapter primer (GTCCAGTCTCGCA GATGATAAGG), 1 µM transposon primer 1 (CCCGAAAAGTGCCACCTAAATTG TAAGCG) or primer 2 (CGCTTCCTCGTGCTTTACGGTATCG), SYBR Green I nucleic acid gel stain (used at manufacturer's recommended concentration), AmpliTaq Gold with GeneAmp PCR Gold buffer, and 1 µL ligated DNA (from **Section 3.4.1**, **step 6**). Use a real-time thermal cycling system (i.e., DNA Engine Opticon 2; Bio-rad, Hercules, CA) to amplify DNA using the following conditions: 95°C for 10 min; 5 cycles of 94°C for 30 s, 72°C or 69°C for 30 s, and 72°C for 1.5 min; 5 cycles of 94°C for 30 s, 70°C or 67°C for 30 s, and 70°C for 1.5 min; 15 to 25 cycles of 94°C for 30 s, 68°C or 65°C for 30 s, and 68°C for 1.5 min + 5 s per cycle; and 72°C for 5 min. The low and high annealing temperatures should be used for primers 1 and 2, respectively. PCR reactions should be removed from the thermocycler during mid-log phase amplification as indicated by fluorescence (**Note 9**).
2. Run the entire PCR reaction on a 2% agarose gel (**Fig. 2**). Successful amplification should give a homogenous smear from approximately 100 to 1000 bp (**Note 10**).
3. Cut out PCR products between 250 and 500 bp and purify them using the QIAquick Gel Extraction Kit (**Note 11**). Wash the DNA bound to the column once with 500 µL QG buffer (supplied with the kit) to remove all traces of agarose and twice with 700 µL PE buffer (supplied with the kit) to prevent carryover of the QG buffer. Contamination of the eluate with agarose or QG can reduce the efficiency of later transcription steps. Elute in 50 µL 2 mM Tris-HCl, pH 8.0.
4. Quantify the products using a fluorometer. Each sample should have at least 2 ng/µL DNA. Mix equal amounts of PCR product amplified with primers 1 and 2 for each sample.
5. Dry the PCR product in a vacufuge.

3.4.3. In vitro *Transcription*

1. Set up *in vitro* transcription reactions using the MEGAShortScript kit. Mix T7 RNA polymerase reaction buffer, 7.5 mM each NTP, 40 U RNaseOUT, 1 µL T7 enzyme mix, and water with the dried PCR product prepared in **Section 3.4.2** in a total volume of 10 µL. Incubate the reaction at 37°C overnight.
2. Digest the DNA template by adding 10 µL water and 1 µL DNase (supplied in the MEGAShortScript kit). Incubate at 37°C for 20 min.
3. Purify RNA by using the RNeasy mini kit. Quantify RNA using a fluorometer or spectrophotometer. Each sample should have at least 5 µg RNA.
4. Mix 10% of the purified RNA with an equal volume of RNA loading buffer (supplied in the MEGAShortScript kit). Heat to 65°C for 3 min. Load the mixture on a 2% agarose gel in TAE buffer. There should be a strong smear between 100 and 400 bp, as indicated by dsDNA standards (**Fig. 3**).
5. Concentrate RNA by evaporation in a vacufuge at 45°C, decreasing the volume to 11 µL. If not used immediately, store at –80°C.

3.4.4. *Synthesis and Labeling of RNA*

1. For each sample, mix RNA from **Section 3.4.3** with 25 µM adapter primer and first strand buffer in a total volume of 20 µL. Heat this mixture to 70°C for 10 min and 42°C for 5 min and then place on ice for at least 1 min.
2. Add 10 µL containing first strand buffer, 15 mM DTT, 3 µL 10 × aa-dNTP mix, 40 U RNase OUT, and 200 U SuperScript III RT enzyme. Incubate at 50°C for 2 h or overnight.
3. To hydrolyze RNA, add 10 µL 0.5 M EDTA and 10 µL 1 M NaOH. Incubate at 65°C for 15 min. Add 10 µL of 1 M HCl to neutralize pH.

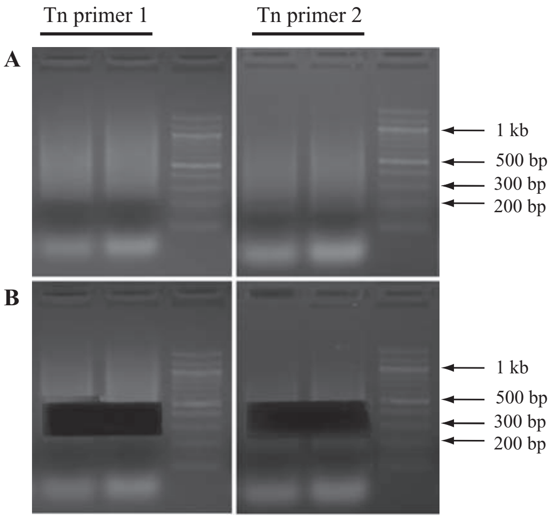


Fig. 2. Transposon-specific PCR. Genomic fragments digested with restriction enzymes and ligated to adaptors were used as PCR template. Adaptor- and transposon-specific primers were used to amplify transposon ends. **(A)** PCR products from two representative samples were resolved on a 2% agarose gel. **(B)** Fragments between 250 and 500 bp in size were extracted and purified.

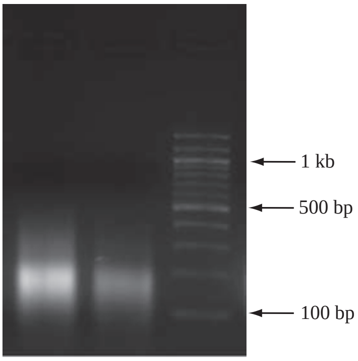


Fig. 3. T7 transcription products. Transposon ends were amplified by PCR, and T7 RNA polymerase was used to transcribe RNA from PCR products using outward-facing T7 promoters at the transposon ends. RNA was purified, and aliquots from two representative samples were loaded on a 2% agarose gel.

4. Purify cDNA using QIAquick PCR purification kit using the following modified steps (**Note 12**). Mix cDNA synthesis products from **step 3** with 20 µL 3 M sodium acetate (pH 5.2) and 350 µL Buffer PB (supplied by Qiagen). Spin the mixture through a QIAquick column. Wash twice with 720 µL cDNA wash buffer. Spin the column dry and then elute in 30 µL water twice.
5. Use a vacufuge to completely dry the eluted cDNA, then resuspend it in 9 µL 0.1 M $NaHCO_3$ at pH 9.0 (**Note 13**).
6. Add to an aliquot of Cy3 or Cy5 monoreactive dye according to manufacturer's instructions. Leave at room temperature for 1 h in the dark.
7. Remove uncoupled dye using QIAquick PCR Purification Kit. Add 35 µL 100 mM sodium acetate (pH 5.2) and 200 µL Buffer PB (supplied by Qiagen) to the labeling reaction. Spin through a QIAquick column. Wash twice with Buffer PE (supplied by Qiagen) and elute in 30 µL 10 mM Tris-HCl, pH 8.0, twice.
8. Use a spectrophotometer to measure yield and labeling efficiency (**Note 14**).

3.5. Microarray Construction, Hybridization, and Analysis

The last steps of TraSH rely on microarray analysis of the regions adjacent to each transposon insertion amplified in the previous sections. Described below are the considerations involved in designing a microarray for TraSH and the steps involved in constructing such an array. Steps involved in microarray hybridization and image acquisition are also described along with statistical considerations involved in data analysis.

3.5.1. Microarray Design

The considerations for designing a microarray for use in TraSH vary considerably from those used for expression analysis, although arrays need not be dedicated to only one application. The main variables in array design are the length of the DNA probe to be immobilized and its position relative to the open reading frame (ORF) it is designed to detect. Neither of these features is critical for the performance of arrays designed to measure mRNA abundance, however, both can be manipulated to optimize

TraSH data. In general, each probe should be complementary to a region near the center of the target ORF. This decreases the probability of detecting insertions in intergenic regions adjacent to the ORF or in nearby ORFs, which are less likely to affect the function of the target ORF. Determining the optimal length of each probe is a trade-off between longer probes, which maximize the number of insertions that are detected, and shorter probes, which are optimal for excluding nondisruptive insertions. In practice, double-stranded probes that are 300 to 500 bp in length work well for most genes, but probes as small as 70 bp have been used successfully (**Note 15**). The following method can be used to design primers specific for each ORF:

1. Use PRIMER3 software *(13)* to generate 10 oligonucleotide primer pairs for each predicted ORF in the *M. tuberculosis* H37Rv genome.
2. Compare each predicted ORF with all others using BLAST and generate a list of any sequence that has an E value of < 1 × 10^{-5} to find ORFs that might misprime. Exclude primers that might anneal to genes in this list.
3. Use BLAST to identify primer pairs that produce fragments that could cross-hybridize to other ORFs (>77% identity) and eliminate these primer pairs.
4. Add 5′ extensions to each forward and reverse primer: GGCATCTAGAG and CCGCACTAGTCCTC, respectively.

3.5.2. Microarray Construction

1. Set up 50-μL PCR reactions with each gene-specific primer pair, 2.5 mM $MgCl_2$, 10% DMSO, 1.25 U Taq, 0.15 U Pfu polymerase, and H37Rv genomic DNA as template. Use the following thermocycling conditions: 94°C for 2 min; 30 cycles of 94°C for 30 s, 60°C for 30 s, 72°C for 1 min; and 72°C for 5 min.
2. Dilute PCR products 1 : 100 and use 2.5 μL in a second-round reaction using similar components, but with universal primers containing 5′ amino modification including a 3-carbon linker (GAACCGATAGGCATCTAGAG and GAAATCCACCGCACTAGTCCTC; IDT). Use the following thermocycling conditions: 95°C for 2 min; 3 cycles of 94°C for 30 s, 40°C for 30 s, and 72°C for 1 min; 20 cycles of 94°C for 30 s, 60°C for 30 s, and 72°C for 1 min; and 72°C for 5 min.
3. Run small aliquots of each second-round reaction on a 2% agarose gel to make sure each primer pair produces a single fragment of the expected size.
4. Purify PCR products using multiscreen PCR plates.
5. Array PCR products onto CodeLink activated slides in duplicate, as recommended by the manufacturer.

3.5.3. Microarray Hybridization

There are several adequate protocols for microarray hybridization. In the past, the authors have successfully used manual hybridization and washing protocols *(2)*. However, we have found more consistent results with lower levels of background using the Tecan HS400 hybridization station. The following describe conditions that can be used with this system.

1. Wash slides printed and processed according to manufacturer's instructions with a solution containing 5× SSC and 0.1% SDS at 42°C for 30 s and allow the slides to soak in the wash for 30 s.

2. Inject 100 μL prehybridization buffer at 42°C. Hybridize at 42°C for 30 min with 1.5 min agitation every 7 min.
3. Wash twice at 23°C with water for 30 s, soaking the slides for 30 s each time.
4. Wash with 5× SSC and 0.1% SDS for 30 s and soak for 30 s, both at 23°C.
5. Inject 50 pmol dye of each labeled cDNA sample suspended in hybridization buffer at 60°C. Heat the slides to 95°C for 2 min to denature the single-stranded cDNA. Hybridize at 42°C for 16 h with 1.5 min agitation every 7 min.
6. Wash with 5× SSC and 0.1% SDS for 30 s and soak for 30 s, both at 23°C.
7. Wash with 0.2× SSC and 0.1% SDS for 30 s and soak for 30 s, both at 23°C.
8. Wash five times with 0.2× SSC for 30 s, soaking for 30 s each time, both at 23°C.
9. Wash with 0.05× SSC for 30 s and soak for 30 s, both at 23°C.
10. Dry slide at 30°C for 90 min.

3.5.4. Image Acquisition and Quantification

The relative amount of each fluorophore that is bound to each probe is quantified using a commercial confocal microarray scanner. Image-quantification software is available commercially or free of charge from Dr. Michael Eisen's lab (http://rana.lbl.gov/). All these programs include functions to assist in the identification of the DNA features spotted on the array and the quantification of the relative amount of each fluorophore bound per spot. In addition, all allow simple data transformations, which will be discussed below.

3.5.5. Statistical Analysis

Replicate TraSH experiments are essential for the statistical analysis of the resulting data. Variability is introduced into TraSH data at multiple points in the procedure beginning with the plating of the library and including the amplification and labeling of the genomic fragments. Therefore, under ideal circumstances, four biological replicates (independently selected and plated libraries) should be analyzed. Analyzing multiple samples generated from the same library (technical replicates) is also useful if multiple libraries are not available. The authors generally perform two technical replicates of 4 to 5 independently plated libraries, and therefore, 8 to 10 microarray hybridizations are available for each analysis. In our experience, the technical replicates tend to be more reproducible than the biological replicates, making the added value of additional biological replicates higher than that of additional technical replicates.

Three data transformations are applied to the raw data that is collected by the microarray scanner. First, the local background intensity is subtracted from each spot. Second, Cy3/Cy5 ratios that are less than 0.01 are set to 0.01, and those that are more than 100 are set to 100. This is based on the assumption that microarray features are unreliable when the values are below or above a certain threshold. After subtraction of background, a feature with little or no fluorescence can have an extremely low or even negative value, creating artifacts that skew the resulting ratio. Third, because the relative intensity of each fluorophore varies from array to array, the data from each array is normalized to ensure that each data set is comparable. Most microarray data is normalized such that the median intensity for each color is equal between arrays. The authors have used LOWESS normalization for this purpose ***(14)*** (*see* http://www.stat.

berkeley.edu/users/terry/zarray/Html/normspie.html). Although most data generated by TraSH appears to be approximately normal, this transformation is problematic for data whose ratios are not normally distributed around 1. Therefore, an alternative normalization strategy that centers the mode of the ratios at 1 should be used for these experiments. One strategy that has been used successfully is to define a set of ~50 genes that are invariably found near the mode of the ratios. The entire data set is then normalized such that the average of these 50 genes equals 1 in all experiments. The ratios for each individual gene can then be averaged across the replicate experiments.

The goal of the statistical analysis of TraSH data is the identification of genes whose ratios are significantly different from 1. Because replicate experiments are performed and, therefore, the distribution of the ratio measurements for each gene is known, a simple *t*-test statistic is adequate for defining these genes. However, more complex statistical tests, which account for increased variance at low intensities, are also valid and may be useful in certain cases, especially when the number of replicates is limited.

In addition to a cutoff based on statistical confidence, an absolute fold–change cutoff is also useful to exclude variability based on the insertional specificity of the transposon. As described in **Section 3.2.1**, the authors have used TraSH analysis to identify genes important for *in vitro* growth by comparing library grown on 7H10 to labeled genomic DNA ***(7)***. In this experiment, cloned DNA fragments were also mutagenized with a *Himar1* transposon and compared with labeled genomic DNA. This resulted in a normal distribution of ratios that varied from 5 to 0.2. Because there is no selection for or against any particular insertion in this case, this variability is due to the insertional specificity of the transposon. Experiments that compare labeled cDNA made from transposon library to labeled genomic DNA must take this insertional bias into account. By identifying genes with ratios that differ from 1 by more than fivefold, variation due to transposon specificity can be largely eliminated, allowing the specific definition of mutants that are underrepresented due to decreased growth rate. More carefully controlled comparisons of transposon libraries selected under one condition to those selected under another condition inherently accounts for insertional bias, allowing the use of less-stringent fold-change cutoffs ***(5, 6)***.

Notes

1. Chloroform, methanol, and phenol are toxic and should be handled accordingly.
2. Although Tween-80 is often used to reduce clumping in mycobacterial cultures, it can also inactivate phage and should therefore be avoided during phage propagation.
3. *M. smegmatis* can typically be grown on LB, but 7H10 agar should be used when preparing a phage stock that will later be used to transduce *M. tuberculosis* or *M. bovis* BCG.
4. All steps that involve live *M. tuberculosis* should be performed under biosafety level 3 conditions.
5. We have found that smaller glass beads tend to stick to the lids of the plates, making them more difficult to remove safely and conveniently.
6. Consistent digestion and enzyme dilution requires well-dissolved genomic DNA. For this reason, we usually dilute 100 μL genomic DNA from **step 9** in **Section 3.3** in 900 μL water and incubate this diluted DNA at 37°C overnight. We have found that using PCR strip tubes and a multichannel pipettor works well for making serial dilutions.

7. HinP1I and MspI were chosen because they cut the GC-rich *M. tuberculosis* genome frequently and create similar overhangs. We are currently working on adapting this method to organisms that have AT-rich genomes using restriction enzymes that have AT-rich recognition sites (i.e., MseI and Tsp509I) and modified adaptors that recognize the corresponding overhangs. When adapting the TraSH technique to other organisms, it is important to use restriction enzymes that are common in the target genome.
8. Be careful to avoid prolonged exposure to UV while cutting DNA samples out of the agarose gel. This is especially important if several samples are being processed at the same time. In this case, only one sample at a time should be exposed to the UV light. Prolonged UV exposure decreases the efficiency of subsequent PCR steps.
9. DMSO is added to the reaction to increase the efficiency of amplification of the GC-rich template DNA.

   The decreasing annealing temperatures used for PCR amplification should minimize amplification due to false priming. This is also the purpose of the Amplitaq Gold enzyme, which becomes activated only after the first 95°C step. As the two primers have different annealing temperatures, we usually use a gradient setting on the thermocycler to allow cycling of both temperatures simultaneously.

   We have found that the most consistent and reliable microarray results are obtained when the lowest number of cycles possible is used to amplify the transposon ends. SYBR Green fluorescent dye is used to indicate the stage of amplification. We have empirically determined a minimum threshold value above the background (as determined by initial fluorescence) that is the lowest value consistently exceeded by early amplification. By removing reactions at several rounds after this value and running the products on an agarose gel, we have found the lowest number of cycles required to consistently provide enough DNA for subsequent steps (about 100 ng after gel extraction). We typically amplify each sample in triplicate and combine triplicate reactions before gel-purification. This allows a higher DNA yield with fewer cycles of amplification. For our thermocycler, we remove reactions six cycles after the fluorescence exceeds 0.09 arbitrary units above the background. This cutoff will need to be empirically determined for each thermocycler used.

   If a real-time thermocycler is not available, replicate PCR reactions can be removed at several different cycles of amplification. Small aliquots of each reaction from different cycles can then be run on an agarose gel. The first reaction that contains a clear smear is used for subsequent steps.

   When several samples are done at the same time, it is helpful to add equal amounts of ligated DNA to each reaction as determined by **step 4**, **Section 3.4.1**. When this is done, the number of cycles needed to reach mid-log phase amplification is more consistent between samples.
10. Poor DNA quality, either as a result of poor ligation efficiency or due to UV damage during gel extraction, can lead to the development of PCR artifacts as indicated by distinct bands visible when the PCR products are run on an agarose gel. We have occasionally found that a band less than 250 bp can occur due to primer 2. As PCR products between 250 and 500 bp are used for the following steps, we usually disregard these artifacts if there is still a strong smear above them. However, distinct bands within the 250 to 500 bp range are cause for concern. Improving the quality of the starting DNA or optimization of the PCR conditions may be required to reduce such artifacts.
11. Care should be taken to avoid RNase contamination of the PCR products prepared in this section and the RNA products prepared in the following section.

12. PE buffer contains Tris, which will inhibit the subsequent NHS coupling reaction. For this reason, we substitute the high-salt cDNA wash buffer.
13. Sodium carbonate ($NaHCO_3$) is made as 1 M solution that is stored at −20°C in multiple aliquots. Solutions with 0.1 M $NaHCO_3$ are made from a single, freshly thawed aliquot of the stock solution. Aliquots are only thawed once to ensure proper pH.
14. The labeling efficiency of the reaction can be calculated as follows: Total cDNA (ng) = $A_{260} \times 37 \times$ volume (μL); Cy3 (pmoles) = $A_{550} \times$ volume (μL)/0.15; Cy5 (pmoles) = $A_{650} \times$ volume (μL)/0.25; nucleotides/dye ratio = Total cDNA (ng) × 1000/(324.5 × pmoles dye). This reaction should yield at least 1 nmol cDNA with one dye molecule per 60 nucleotides.
15. In the process of preparing this manuscript, there have been preliminary reports of the use of 70-mer oligonucleotide arrays for TraSH analysis in *M. tuberculosis*. These arrays, made available through The Institute for Genomic Research, have one 70-mer oligonucleotide for each predicted ORF. Although there is no reason that this type of array should not be used, the utility of this system has not been rigorously demonstrated to date.

    Alternative oligonucleotide array designs use multiple small oligonucleotides for each ORF as well as intergenic regions (originally developed by Affymetrix). Such a design offers advantages in that it could be used to identify subgenic regions associated with specific phenotypes. It also allows the identification of phenotypes associated with unannotated regions of the genome such as small ORFs or RNAs. Unfortunately, technical considerations make it difficult for individual labs to produce such arrays, and commercial products are not yet available for many organisms.

Acknowledgments

The authors thank David Alland and Hassan Safi for their helpful suggestions. This work was supported by NIH grants AI48704 and AI51929 and by the Burroughs Wellcome Foundation (E.J.R.).

References

1. Badarinarayana, V., Estep, P. W. 3rd, Shendure, J., Edwards, J., Tavazoie, S., Lam, F., and Church, G. M. (2001) Selection analyses of insertional mutants using subgenic-resolution arrays. *Nat. Biotechnol.* **19**, 1060–1065.
2. Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2001) Comprehensive identification of conditionally essential genes in mycobacteria. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 12,712–12,717.
3. Salama, N. R., Shepherd, B., and Falkow, S. (2004) Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori*. *J. Bacteriol.* **186**, 7926–7935.
4. Chan, K., Kim, C. C., and Falkow, S. (2005) Microarray-based detection of *Salmonella enterica* serovar Typhimurium transposon mutants that cannot survive in macrophages and mice. *Infect. Immun.* **73**, 5438–5449.
5. Rengarajan, J., Bloom, B. R., and Rubin, E. J. (2005) Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 8327–8332.
6. Sassetti, C. M., and Rubin, E. J. (2003) Genetic requirements for mycobacterial survival during infection. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12,989–12,994.
7. Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* **48**, 77–84.

8. Lawley, T. D., Chan, K., Thompson, L. J., Kim, C. C., Govoni, G. R., and Monack, D. M. (2006) Genome-wide screen for salmonella genes required for long-term systemic infection of the mouse. *PLoS Pathog.* **2**, e11.
9. Bardarov, S., Kriakov, J., Carriere, C., Yu, S., Vaamonde, C., McAdam, R. A., et al. (1997) Conditionally replicating mycobacteriophages: a system for transposon delivery to *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 10,961–10,966.
10. Lampe, D. J., Akerley, B. J., Rubin, E. J., Mekalanos, J. J., and Robertson, H. M. (1999) Hyperactive transposase mutants of the *Himar1 mariner* transposon. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 11,428–11,433.
11. Rubin, E. J., Akerley, B. J., Novik, V. N., Lampe, D. J., Husson, R. N., and Mekalanos, J. J. (1999) *In vivo* transposition of *mariner*-based elements in enteric bacteria and mycobacteria. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 1645–1650.
12. Belisle, J. T., and Sonnenberg, M. G. (1998) Isolation of genomic DNA from mycobacteria. *Methods Mol. Biol.* **101**, 31–44.
13. Rozen, S., and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–386.
14. Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001) Normalization for cDNA microarray data. Presented at SPIE BiOS 2001, San Jose, California.

5

Essential Genes in the Infection Model of *Pseudomonas aeruginosa* PCR-Based Signature-Tagged Mutagenesis

François Sanschagrin, Irena Kukavica-Ibrulj, and Roger C. Levesque

Summary

PCR-based signature tagged mutagenesis is an "en masse" screening technique based upon unique oligonucleotide tags (molecular barcodes) for identification of genes that will diminish or enhance maintenance of an organism in a specific ecological niche or environment. PCR-based STM applied to *Pseudomonas aeruginosa* permitted the identification of genes essential or *in vivo* maintenance by transposon insertion and negative selection in a mixed population of bacterial mutants. The innovative adaptations and refinement of the technology presented here with *P. aeruginosa* STM mutants selected in the rat lung have given critical information about genes essential for causing a chronic infection and a wealth of information about biological processes *in vivo*. The additional use of competitive index analysis for measurement of the level of virulence *in vivo*, microarray-based screening of selected prioritized STM mutants coupled to metabolomics analysis can now be attempted systematically on a genomic scale. PCR-based STM and combined whole-genome methods can also be applied to any organism having selectable phenotypes for screening.

Key Words: attenuation of virulence; competitive index; en masse screening; signature-tagged mutagenesis.

1. Introduction

A combination of bacterial and molecular genetic techniques, the so-called genomics-based technologies, can now be used to study bacterial pathogenesis on a global scale at the genome level and *in vivo* (*1, 2*). These methods include *in vivo* expression technology, or IVET (*3*) (promoter trap for genes expressed solely *in vivo*), DNA chips (transcriptomics profiling), proteomics (via differential display in 2D gels), differential hybridization (selective expression *in vitro* vs. *in vivo* of specific transcripts), and signature-tagged mutagenesis (STM) (based on phenotypic attenuation of virulence).

Of these methodologies, STM is of particular interest. This elegant bacterial genetics method is based on negative selection to identify mutations in genes that are essential during the infection process (*4, 5*). In STM, transposon mutants are generated, and each

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

unique bacterial clone is tagged with a specific DNA sequence that can be rapidly identified by hybridization or more easily by PCR in a pool of mutants. STM is an "en masse" screening technique where a tagged mutant having an insertion in a gene causing a defect in virulence will be out-competed. It minimizes the number of animals used by pooling mutants. In this negative selection scheme, the mutant bacteria cannot be maintained *in vivo*; technically, attenuated mutants are selected by the host and identified by comparing the *in vitro* input and the *in vivo* output pools of mutants using multiplex polymerase chain reaction (PCR). STM mutants identified are retested to confirm attenuation in virulence when compared with the wild-type strain; disrupted genes are cloned via the transposon marker, and the inactivated genes are identified by DNA sequencing.

Recent modifications of STM to eliminate the hybridization steps allow rapid and easy identification of attenuated mutants using multiplex and real-time PCR. We refer to this method as PCR-based STM (***2, 6, 7***). This PCR-based STM is an extremely powerful and elegant bacterial genetics approach for *in vivo* functional genomics, particularly when used in combination with bioinformatics, proteomics, transcriptomics, and metabolomics analysis to identity genes and their products essential for *in vivo* maintenance (***8***).

As an example of PCR-based STM, we will use the opportunistic pathogen *Pseudomonas aeruginosa*, which has the remarkable ability to adapt to various ecological niches. The 6.3-Mb genome of *P. aeruginosa* strain PAO1 has been completely sequenced, and its annotation is available at: http://www.pseudomonas.com (***9, 10***). The sequence of strain PAO1 is of particular interest for STM analysis (***11***) because it encodes 5570 open reading frames (ORFs), which comprises more than 543 regulatory motifs characteristic of transcriptional regulators, 55 sensors, 89 response regulators, and 14 sensor–response regulatory hybrids of two-component systems and at least 12 potential resistance-nodulation-cell division (RND) efflux systems including 300 proteins implicated in transport (65% would be implicated in nutrient uptake). Because more than 45% of ORFs from the sequence of PAO1 contained hypothetical proteins, we felt that this was a gold mine for identifying particular virulence factors of opportunistic pathogens and genes essential for *in vivo* maintenance. As summarized in **Figure 1**, the functions of most proteins encoded by the *P. aeruginosa* PAO1 genome are barely known, and PCR-based STM is a powerful tool for this analysis. The PAO1 genome encodes 1780 (32%) genes having no homology to any previously reported sequences; 1590 (28.5%) genes having a function proposed based on the presence of conserved amino acid motif, structural features, or limited homology; and 769 (13.8%) homologues of previously reported genes of unknown function. In terms of genes characterized, 1059 (19%) have a function based on a strongly homologous gene experimentally demonstrated in another organism, whereas only 372 genes (6.7%) have a function experimentally demonstrated in *P. aeruginosa*.

The PCR-based STM method (***12***) has been applied extensively to *P. aeruginosa* PAO1 and will be used to illustrate the methods utilized for construction of the mutant libraries, the preparation of agar beads for *in vivo* screening in a rat model of chronic lung infection, the identification of mutants by multiplex PCR, the selection of mutants attenuated for *in vivo* maintenance, and their analysis using a competitive index.

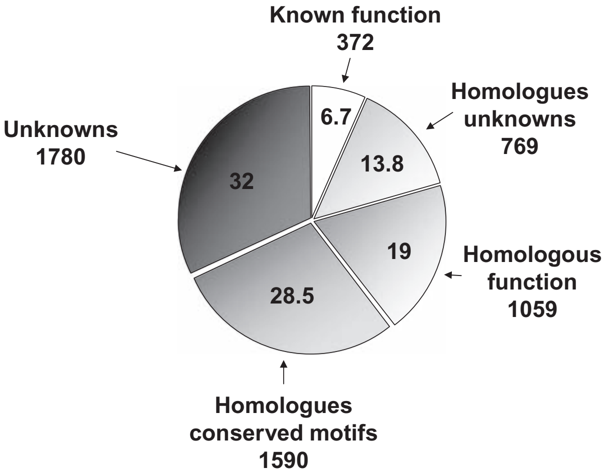


Fig. 1. General features of the 5570 ORFs from *P. aeruginosa* (*9*). The number of ORFs in each of the five groups of protein coding sequences—known function, homologous unknown, homologous function, homologues conserved motifs, and unknowns—is indicated. The percentage represented by each ORF group is indicated.

2. Materials

1. Plasmids: pUTmini-Tn*5* Km2, pUTmini-Tn*5* Tc (*13*), pUTmini-Tn*5* TcGFP (*14*), pTZ18R (GE Healthcare, Baie d'Urfé, Québec, Canada), pPS856, pDONR221, pEX18ApGw (*15*), pUCP19 (*16*).
2. Oligonucleotides for tag construction and universal primers for multiplex PCR listed in **Table 1**.
3. 10× medium salt buffer (oligonucleotide buffer): 10 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$, 50 mM NaCl, 1 mM DTT.
4. Deoxynucleotide triphosphates (dNTPs): dATP, dGTP, dCTP, dTTP.
5. Restriction enzymes: T4 DNA polymerase, HotStartTaq DNA polymerase (Qiagen, Mississauga, Ontario, Canada); T4 DNA ligase and HiFi Platinum Taq (Invitrogen, Burlington, ON, Canada).
6. Restriction enzymes buffers: 10× NEB 1, 2, 3, (New England Biolabs [NEB], Mississauga, ON, Canada).
7. 10× BSA (1 mg/mL) (NEB).
8. T4 DNA ligase 10× buffer (NEB).
9. Micropure-EZ pure, microcon 30, microcon PCR (Millipore, Nepean, ON, Canada).
10. *P. aeruginosa* strain PAO1 (*17*).
11. *Escherichia coli* strains, S17–1 λ *pir*, DH5α, ElectroMax DH10B (Invitrogen), One Shot MAX Efficiency DH5α-T1^r (Invitrogen).
12. Bio-Rad GenePulser.
13. Electroporation gap cuvettes, 1 mm and 2 mm.
14. Hotplate stirrer (Corning, Model 4200, Fisher Scientific, Québec, Canada).
15. Bacterial growth media: tryptic soy broth (TSB), brain heart infusion, (BHI), tryptic soy agar (TSA), Mueller-Hinton agar (MHA), *Pseudomonas* isolation agar (PIA), BHI agar.
16. Antibiotics: ampicillin (Ap), kanamycin (Km), tetracycline (Tc), gentamicin (Gm), carbenicillin (Cb), chloramphenicol (Cm).
17. TE PCR buffer: 10 mM Tris-HCl pH 7.4; EDTA 0.1 mM.

18. 10× HotStartTaq DNA polymerase reaction buffer with Tris-Cl, KCl, $(NH_4)_2SO_4$, 15 mM $MgCl_2$, pH 8.7 (Qiagen).
19. 10 pmol oligonucleotide tags, universal primers listed in **Table 1**.
20. Mineral oil.
21. Agarose LM Nusieve GTG (FMC, Rockland, Maine).
22. Standard gel electrophoresis grade agarose, 1× Tris-borate EDTA buffer, and 0.5 μg/mL ethidium bromide solution.
23. MF-Millipore membrane filter 0.025 μm, 25 mm (Millipore).
24. Sterile 1× phosphate-buffered saline (PBS): 137 mM NaCl, 3 mM KCl, 10 mM Na_2HPO_4, 1.3 mM KH_2PO_4 pH 7.4.
25. 2-mL 96-well plates (Qiagen).
26. Sprague-Dawley rats, 450 to 500 g, male.
27. Polytron homogenizer (Kinematica AG, Lucerne, Switzerland).
28. QIAGEN Dneasy Tissue kit (Qiagen).
29. QIAfilter plasmid midi kit (Qiagen).
30. QIAquick gel extraction kit (Qiagen).
31. Quant-iT Picogreen dsDNA reagent and kit (Invitrogen).
32. Gateway BP Clonase II Enzyme Mix (Invitrogen).
33. Gateway LR Clonase II Enzyme Mix (Invitrogen).
34. DNA sequencing service and bioinformatics software.

3. Methods

STM is divided into two major steps: the construction of a library of tagged mutants by transposon mutagenesis, which implicates the synthesis and ligation of DNA tags into a specific site, transfer of the transposon into the recipient host, selection of transconjugants, and arraying of the mutants; and the *in vivo* screening step, which involves an *in vivo* animal or cell model of selection, the screening of tissues for mutant bacteria, and comparative PCR analysis of mutants not found in the host because STM is a negative selection process (*2*). A crucial step in STM depends upon a high frequency of random transposon insertions into the chromosome. This is not always possible because of low frequencies of transposition in certain bacterial hosts and the presence of hot spots of insertion in certain bacterial genomes. When applying STM, one must take into consideration that insertion into an essential gene gives a lethal phenotype. These genes cannot be identified by STM, and several may be critical for virulence (*12*). Obviously, STM will identify only mutants attenuated for *in vivo* maintenance when compared with the wild-type strain used. All mutants selected require several rounds of *in vivo* screening, testing for auxotrophy, and analysis by a competitive index (CI) to estimate changes in the level of virulence for a particular mutant when compared with the wild type.

The methods below outline the construction of tagged plasmids including tag annealing, plasmid preparation, plasmid and tag ligation, and electroporation (**Section 3.1**); construction of libraries of tagged mutants by conjugation including transposon mutagenesis (**Section 3.2**); *in vivo* screening of tagged mutants insertion of *P. aeruginosa* into agar beads to facilitate initiation of a chronic infection in the rat lung for the first *in vivo* passage of tagged mutants (**Section 3.3**); cloning, sequencing, and analysis of disrupted genes responsible for attenuation of virulence in STM mutants (**Section 3.4**); construction of gene knockouts for selected STM mutants (**Section 3.5**); and a

Table 1
Nucleotide Sequences of the 24 Oligonucleotides Used for Construction of Signature Tags and Sequences of the Three Universal Primers for Multiplex PCR-Based STM

Tag Number	Nucleotide sequence
1	GTACCGCGCTTAA**ACGTTCA**G
2	GTACCGCGCTTAA**ATAGCCT**G
3	GTACCGCGCTTAA**AAGTCTC**G
4	GTACCGCGCTTAA**TAACGTG**G
5	GTACCGCGCTTAA**ACTGGTA**G
6	GTACCGCGCTTAA**GCATGTT**G
7	GTACCGCGCTTAA**TGTAACC**G
8	GTACCGCGCTTAA**AATCTCG**G
9	GTACCGCGCTTAA**TAGGCAA**G
10	GTACCGCGCTTAA**CAATCGT**G
11	GTACCGCGCTTAA**TCAAGAC**G
12	GTACCGCGCTTAA**CTAGTAG**G
13	CTTGCGGCGTATT**ACGTTCA**G
14	CTTGCGGCGTATT**ATAGCCT**G
15	CTTGCGGCGTATT**AAGTCTC**G
16	CTTGCGGCGTATT**TAACGTG**G
17	CTTGCGGCGTATT**ACTGGTA**G
18	CTTGCGGCGTATT**GCATGTT**G
19	CTTGCGGCGTATT**TGTAACC**G
20	CTTGCGGCGTATT**AATCTCG**G
21	CTTGCGGCGTATT**TAGGCAA**G
22	CTTGCGGCGTATT**CAATCGT**G
23	CTTGCGGCGTATT**TCAAGAC**G
24	CTTGCGGCGTATT**CTAGTAG**G
pUTKana2	GGCTGGATGATGGGGCGATTC
pUTgfpR2	ATCCATGCCATGTGTAATCCC
tetR1	CCATACCCACGCCGAAACAAG
Gm-F*	CGAATTAGCTTCAAAAGCGCTCTGA
Gm-R*	CGAATTGGGGATCTTGAAGTTCCT
GW-*attB*1*	GGGGACAAGTTTGTACAAAAAAGCAGGCT
GW-*attB*2*	GGGGACCACTTTGTACAAGAAAGCTGGGT
PA2896-UpF-GWL*	TACAAAAAAGCAGGCTcgaaggatgtggccgatgag
PA2896-UpR-Gm*	TCAGAGCGCTTTTGAAGCTAATTCGatcaggctgagccaggtttc
PA2896-DnF-Gm*	AGGAACTTCAAGATCCCCAATTCGacagcgcgaggtattcctg
PA2896-DnR-GWR*	TACAAGAAAGCTGGGTggaaatgcgccagcatctg

Each of 21-mers has a T_m of 64°C and permits PCR amplification in one step when the three primer combinations are used for multiplex screening. Two sets of consensus 5′-ends comprising the first 13 nucleotides have higher ΔG's for optimizing PCR. Twelve variable 3′-ends define tag specificity and allow amplification of specific DNA fragments. The set of twenty-four 21-mers representing the complementary DNA strand in each tag are not represented and can be deduced from the sequences present. Single colonies are selected, purified, and screened by colony PCR using 10 pmol pUTKana2, pUTgfpR2, and tetR1 as the 3′ primers designed in the transposon resistance gene for multiplex PCR. Unique nucleotide sequences in each oligonucleotide primer (1 to 24) is indicated in bold.

*Sequences in capital letters are common for all genes to be replaced and overlap with the Gm or *attB* primer sequences. Lowercase letters indicate gene-specific sequences; here, *PA2896* is used as an example.

competitive index analysis of selected mutants to estimate the level of attenuation of virulence (**Section 3.6**).

3.1. Construction of Tagged Plasmids

The PCR-based STM scheme involves designing pairs (24 in this case, but 48 and 96 unique oligonucleotides could be utilized) of 21-mers (**Table 1**) synthesized as complementary DNA strands for cloning into the mini-Tn*5* plasmid vectors as shown in **Figure 2**. The sets of 24 tags are repeatedly used to construct 24 libraries as shown in **Figure 3A**. DNA amplification using a specific tag as a PCR primer coupled to three primers specific to the Km, Tc, and GFP genes gives three products of specific size easily detectable by multiplex PCR depicted in **Figure 3B**. Multiplex PCR products obtained from arrayed bacterial clones *in vitro* can be compared with the amplified DNA products obtained after *in vivo* passage. These PCR products can easily be visualized in agarose gels as 980-, 820-, and 220-bp amplified products as depicted in **Figure 3B**.

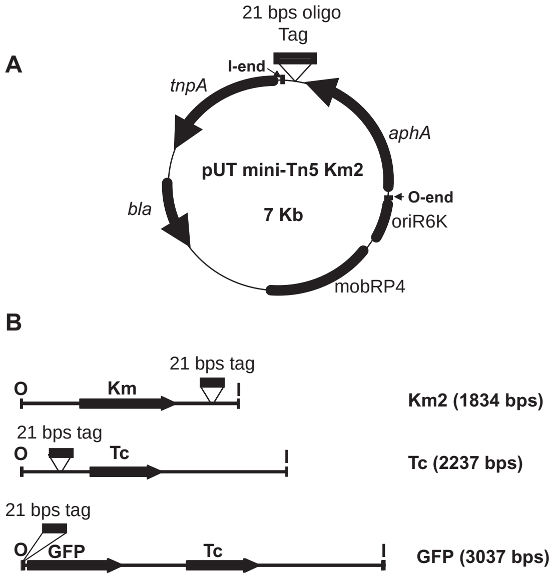


Fig. 2. (**A**) Physical and genetic maps of the pUT plasmid and the mini-Tn*5*Km2, mini-Tn*5*Tc, and mini-Tn*5*GFP transposons used. The transposons are located on an R6K-based suicide delivery plasmid pUT where the Pi protein is furnished by the donor cell (*E. coli* S17-1 λ pir); the pUT plasmid provides the IS50R transposase in *cis*, but the *tnp* gene is external to the mobile element and its conjugal transfer to recipients is mediated by RP4 mobilization functions in the donor (*21*). (**B**) The elements are represented by thick black lines, inverted repeats are indicated as vertical boxes, and genes are indicated by arrows. This collection of Tn5-derived mini-transposons has been constructed that simplifies substantially the generation of insertion mutants, *in vivo* fusions with reporter genes, and the introduction of foreign DNA fragments into the chromosome of a variety of Gram-negative bacteria. The mini-Tn*5* consists of genes specifying resistance to Km, Tc, and GFP with unique cloning sites for tag insertion flanked by 19-base-pair terminal repeats, the I and the O ends. I and O, inverted repeat ends; Km, kanamycin; Tc, tetracycline; GFP, green fluorescent protein.

Instead of complicating PCR analysis using 72 or 96 unique PCR tags, we prepared 24 pairs of 21-oligomers coupled to three distinct phenotypic selections of transposon markers such as Km, Tc, and Tc with green fluorescent protein (GFP) but still giving a total of 72 distinct tags (***7***). We reasoned that a rapid analysis of 24 PCR reactions in multiplex format is more straightforward, rapid, and easier to perform than 72 single PCR reactions.

The oligonucleotides were designed as tags following three basic rules: (a) similar T_m of 64°C to simplify tag comparisons by using one step of PCR reactions; (b) invariable 5′-ends with higher ΔG than at the 3′-end to optimize PCR amplification reactions; (c) a variable 3′-end for an optimized yield of specific amplification product from each tag (***18, 19***). The 21-mers are annealed double-stranded and are cloned into a mini-transposon (mini-Tn*5*), which is used for insertional mutagenesis and, hence, tag bacteria. This collection of transposons can be used with any bacterial system that can conjugate with *E. coli* as a donor and is available upon request.

3.1.1. Tag Annealing

A collection of 24 defined 21-mers oligonucleotides should be synthesized along with their complementary DNA strands using the templates listed in **Table 1**. Annealing reactions contained 50 pmol of both complementary oligonucleotides in 100 µL of 1× medium salt buffer. This oligonucleotide mixture is heated 5 min at 95°C, left to cool slowly at room temperature in a block heater, and kept on ice.

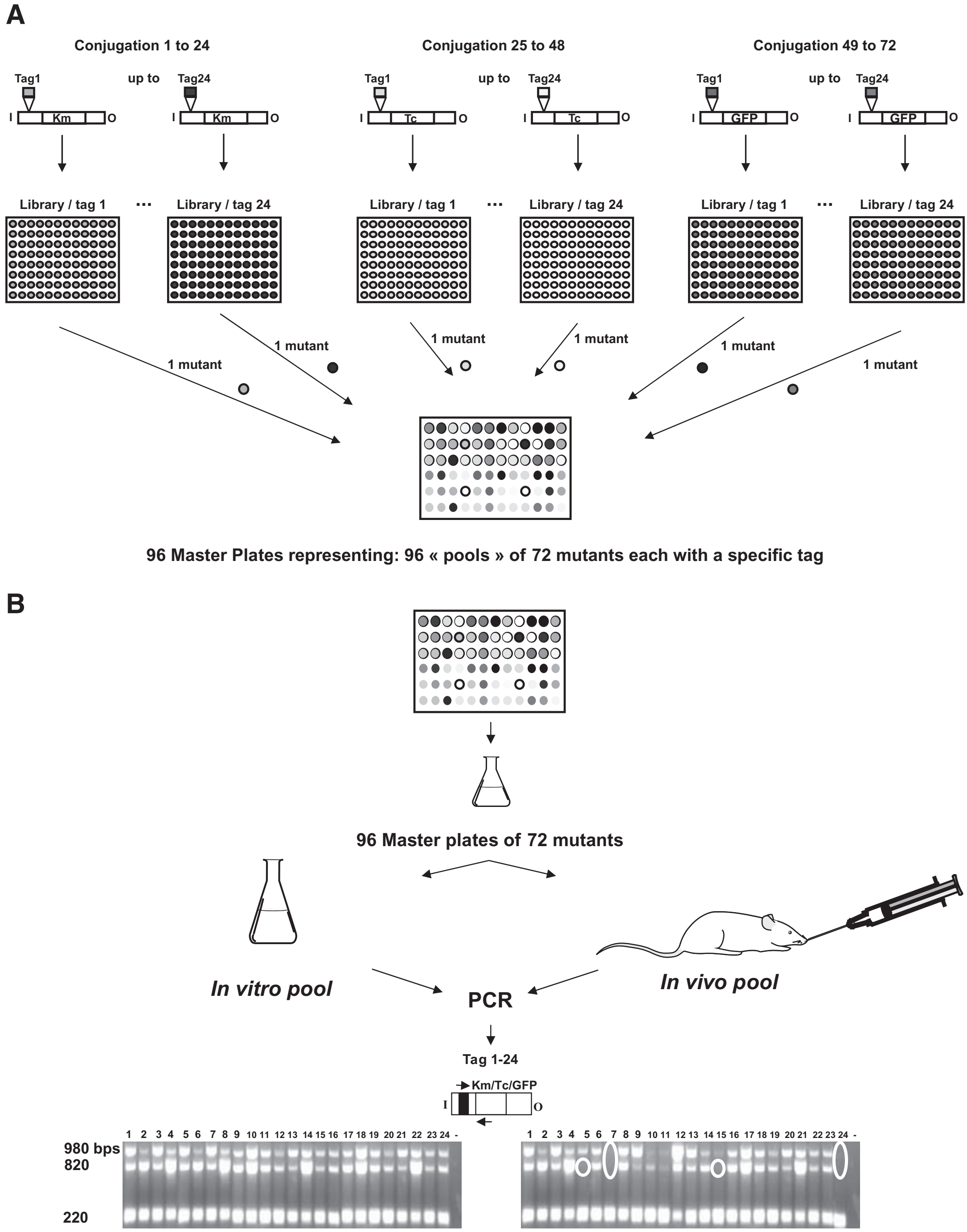
3.1.2. Plasmid Purification and Preparation for Tag Ligation

On a routine basis, we use the Qiagen system for plasmid preparation. DNA manipulations were performed by standard recombinant DNA procedures (***20***).

1. 20 µg of each pUTmini-Tn*5* plasmid DNA is digested with 20 units of *Kpn*I in 40 µL of 1× NEB 1 buffer containing 1× BSA, and incubated for 2 h at 37°C, and the enzyme is inactivated for 20 min at 65°C.
2. Extremities are blunted with T_4 DNA polymerase by adding 4 nmol of each dNTP and 5 units of T_4 DNA polymerase.
3. Purify each blunted plasmid DNA to eliminate endonuclease and T_4 DNA polymerase reactions with micropure-EZ and microcon 30 systems in a single step as described by the manufacturer’s protocol.

3.1.3. Plasmid and Tag Ligation and Electroporation

1. Each plasmid (0.04 pmol) is ligated to 1 pmol of double-stranded DNA tags in a final volume of 10 µL of T_4 DNA ligase 1× buffer containing 400 units of T_4 DNA ligase. Note that 24 ligation reactions are performed for each plasmid, which implies 72 single reactions, 72 electroporations, and 72 PCR analyses.
2. Ligated products are purified using microcon PCR (Millipore) as described by the manufacturer’s instructions and resuspended in 5 µL H_2O.
3. The 5-µL solution containing ligated products are introduced into *E. coli* S17-1 λ pir by electroporation using a Bio-Rad apparatus (2.5 KV, 200 Ohms, 25 µF) in a 2-mm electroporation gap cuvette. After electroporation, 0.8 mL SOC is added to the bacterial preparation, and the solution is transferred in culture tubes for incubation for 1 h at 37°C.


A
Conjugation 1 to 24
Conjugation 25 to 48
Conjugation 49 to 72
Tag1
up to
Tag24
I
Km
O
Tc
GFP
Library / tag 1
Library / tag 24
1 mutant
96 Master Plates representing: 96 « pools » of 72 mutants each with a specific tag
B
96 Master plates of 72 mutants
In vitro pool
In vivo pool
PCR
Tag 1-24
Km/Tc/GFP
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 -
980 bps
820
220

4. Transformed bacteria containing tagged plasmids are selected on TSB supplemented with 50 µg/mL Ap and 50 µg/mL Km by plating 100 µL of electroporated cells.
5. Single colonies are selected, purified, and screened by colony PCR in 50-µL reaction volumes containing 10 µL of boiled bacterial colonies in 100 µL TE PCR buffer; 5 µL of 10× HotStartTaq polymerase reaction buffer; 1.5 mM MgCl2; 200 µM of each dNTPs 10 pmol of one of the oligonucleotides used for tags as a specific 5′ primer and 10 pmol of the pUTKanaR1, the pUTgfpR2, and the tetR1 (**Table 1**) as the universal 3′ primer; and 2.5 units HotStartTaq polymerase (Invitrogen). Thermal cycling conditions are for touchdown PCR including:
   (a) a hot start for 15 min at 95°C;
   (b) 22 cycles at 95°C for 1 min, varied annealing temperature for 1 min (after cycle 2 decrease the temperature from 70°C to 60°C by 1°C every 2 cycles) and at 72°C for 1 min;
   (c) followed by 10 cycles at 95°C for 1 min, 60°C for 1 min, and 72°C for 1 min.
6. Amplified products (10-µL aliquots) are analyzed by electrophoresis in a 1% agarose gel, 1× Tris-borate EDTA buffer, and stained for 10 min in 0.5 µg/mL ethidium bromide solution (***20***) (**Note 1**).

3.2. Construction of Libraries of Tagged Mutants

A series of suicide pUT plasmids carrying mini-Tn*5*Km2, mini-Tn*5*Tc, and mini-Tn*5*Tc-GFP each with a specific tag were transferred by conjugation (***21***) into the targeted bacteria *P. aeruginosa* giving 72 libraries of mutants; 96 mutants each of 72 libraries were arrayed into 96-well master plates (**Fig. 3A, B**). The 72 mutants from the same pool were grown separately overnight at 37°C. Aliquots of these cultures were pooled and a sample kept for PCR analysis (the *in vitro* pool). A second sample from the same pool was used for the *in vivo* passage.

Fig. 3. (**A**) Construction of master plates of *P. aeruginosa* STM mutants for *in vitro* and *in vivo* screening by PCR-based STM. Each master plate contains a collection of 72 mutants having unique chromosomal transposon insertions and are selected from arrayed mutants obtained by conjugation. As depicted above, each conjugation set for a transposon is done using a specific marker (kanamycin, Km; tetracycline, Tc; and Tc-GFP, green fluorescent protein) containing 24 tags. Selection is based on antibiotic-resistance markers and PCR for each set of specific tags. The shading in plates indicates a particular tag; the shading of bacteria in the master plate represents a unique mutant with a transposon insertion. The open-boxed lines represent each transposon, and I and O ends inverted repeats are indicated. The pUTmini-Tn*5*Km, Tc, and GFP vectors were used. (**B**) Comparative analysis between the *in vitro* and *in vivo* pools using multiplex PCR. An aliquot is kept as the *in vitro* pool, and a second aliquot from the same preparation is used for passage into the rat lung for negative selection. At determined time points of infection, bacteria are recovered from the lung and constitute the *in vivo* pool. The *in vitro* and *in vivo* pools are used to prepare DNA in 24 PCR multiplex reactions using the 24 specific 21-mers tags and the Km-, Tc-, and GFP-specific primers. Comparisons between *in vitro* and *in vivo* multiplex PCR products are done by agarose gel electrophoresis for identification of mutants absent *in vivo* (indicated by the white halos in lanes 5, 7, 15, and 24). The PCR products of 980, 820, and 220 bp when amplified with Tc, GFP, Km, and tag-specific PCR primers, respectively. Each mutant is confirmed by a specific PCR; resistance markers are cloned and flanking regions sequenced to identify the inactivated gene.

3.2.1. Conjugation and Transposon Mutagenesis

1. *E. coli* S17-1 λ *pir* containing the pUTmini-Tn*5* tagged plasmids is used as a donor for conjugal transfer into the recipient strain. The ratio of donor-to-recipient bacterial cells to obtain the maximum of exconjugants should be determined in preliminary experiments. For *P. aeruginosa*, we used 1 donor to 10 recipient cells. Cells are mixed and spotted as a 50-μL drop on a membrane filter placed on a nonselective BHIA plate. Plates are incubated at 30°C overnight.
2. Filters are washed with 10 mL of PBS saline to recover bacteria.
3. Aliquots of 100 μL of the PBS solution containing exconjugants are plated on five BHIA plates supplemented with the appropriate antibiotic to select for the strain. For *P. aeruginosa*, we use Km (350 or 500 μg/mL) and Tc (15 or 30 μg/mL). Plates are incubated overnight at 37°C.
4. Km- or Tc-resistant *P. aeruginosa* exconjugants are arrayed as libraries of 96 clones in 2-mL 96-well plates in 1.5 mL of BHI supplemented with Km and appropriate antibiotic. The 2-mL 96-well plates are incubated 18 to 22 h at 37°C (**Note 2**).
5. As an STM working scheme, one mutant from each library is picked to form 96 pools of 72 unique tagged mutants (**Fig. 3A**) contained in the 2-mL 96-well plates.

3.3. In Vivo *Screening of Tagged Mutants*

Unfortunately, traditional screening in animal models of infection for mutants covering a complete genome and based on a gene by gene mutational approach is not feasible *in vivo*, even with today's capabilities in genomics and proteomics. For example, a significant analysis of virulence determinants for the *P. aeruginosa* 6.3-Mb genome encoding 5570 ORFs would require in a model of infection a minimum of 5570 animals; statistical validity would recommend groups of at least five individuals giving a total of 27,850 animals; an impossible and unjustifiable task.

Bacteria are recovered from the lung of each animal (the *in vivo* pool), and the *in vitro* pools are used as templates in 24 distinct multiplex PCR reactions. PCR products are separated by gel electrophoresis where the presence or absence of DNA fragments and their sizes are compared between the *in vitro* and *in vivo* pools. Mutants whose PCR products have not been detected after the *in vivo* passage are *in vivo* attenuated (**Fig. 3B**). This simple STM method can be adapted to any bacterial system and used for genome scanning in various growth conditions.

3.3.1. Preparation of Arrayed Bacteria for In Vitro *PCR*

1. The 72 mutants from the same pool are grown separately overnight at 37°C in 200 μL TSB containing Km or Tc in 96-well microtiter plates.
2. Aliquots of these cultures are pooled.
3. A first sample is diluted from 10^{-1} to 10^{-4} and plated on BHIA supplemented with the appropriate antibiotic for each transposon marker (Km or Tc).
4. After overnight incubation at 37°C, 10^4 colonies are recovered in 5 mL PBS, and a sample of 1 mL is removed for PCR and called the *in vitro* pool.
5. The 1 mL *in vitro* pool sample is spun down, and the cell pellet is resuspended in 1 mL TE PCR buffer.
6. The *in vitro* pool is boiled 10 min and spun down, and 10 μL of supernatant are used in PCR analysis as described above.
7. A second sample from the pooled cultures is used to inoculate animals.

3.3.2. Preparation of Agar Beads with Pools of 72 Mutants

We use two methods for enmeshing *P. aeruginosa* cells into agar beads (*22, 23*). For large-scale library screening of pooled mutants, we use a centrifugation technique (*see* below), and for selected STM mutants in competitive index analysis, we use a decantation technique (**Section 3.6**). Both methods give the same type and yield of agar beads and infection kinetics. The general scheme for agar-bead preparation is given in **Figure 4** (**Note 3**).

3.3.2.1. Day 1

1. Inoculate pool of *P. aeruginosa* STM mutants in 10 mL TSB with appropriated antibiotics in a 250-mL Erlenmeyer flask or in 200 µL of TSB in a deep 96-well plate (TSB + Cm 5 µg/mL for *P. aeruginosa*).
2. Incubate 17 h at 37°C without agitation.
3. Prepare 10 mL PBS containing 2% agar for each bead preparation and sterilize by autoclaving.
4. Prepare and sterilize a large supply of PBS 1×, centrifugation bottles, 200 mL mineral oil in a 250-mL Erlenmeyer flask with a magnetic stirrer, BHIA with and without antibiotics, and feeding needles.

3.3.2.2. Day 2

1. A 2% agar solution is melted in a microwave and separated in 10-mL aliquots for bead preparations in separate culture tubes (13 × 100 mm).
2. Culture tubes and Erlenmeyer flask containing mineral oil are placed in a water bath at 48°C.
3. 0.5 mL of each pooled culture is washed twice with the same volume of PBS 1×, and centrifugations are done at 7200 rpm for 2 min.

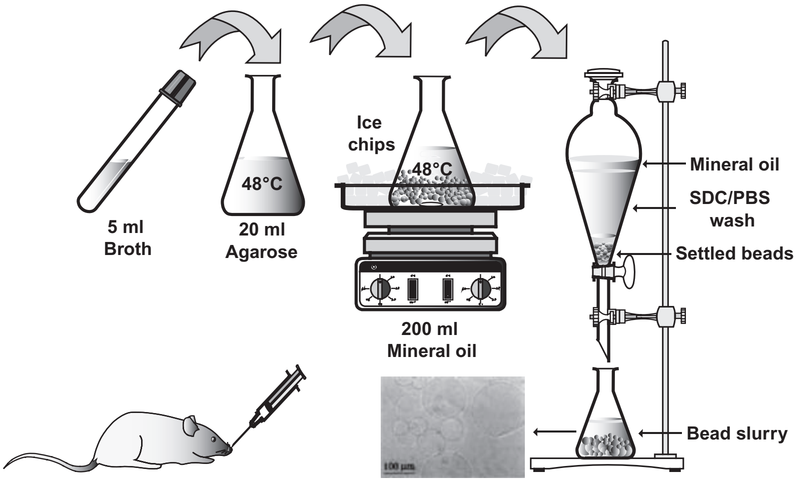


Fig. 4. Preparation of encapsulated *P. aeruginosa* in agar beads. The basic setup is presented using basic microbiological techniques. Beads can be observed with an inverted light microscope using a 10× objective. Details of the preparation steps, determination of colony-forming units prior to infection, and analysis are given in the text (**Section 3.3.2** and Refs. *8, 22, 23*).

4. 50 µL of washed culture is added to 10 mL of 2% agar solution. Vortex the agar bacterial mixture.
5. Place Erlenmeyer flask containing mineral oil into a Pyrex container half-filled with water; place on a magnetic stirrer and start stirring.
6. The agar-bacterial mixture solution is poured into the mineral oil in the center of the vortex (not on the side of the Erlenmeyer flask) while stirring.
7. A mixture of water-ice "slush" is rapidly added on the side of the Erlenmeyer flask to cool the solution in the Pyrex container. Stirring is maintained for 5 min.
8. The agar preparation is placed at room temperature for 10 min without stirring, allowing agar beads to settle at the bottom of the Erlenmeyer flask.
9. A Pasteur pipette hooked to a vacuum is used to remove half of the mineral oil.
10. Agar beads are poured into a 250-mL polycarbonate centrifugation bottle, and the volume is completed to 200 mL with 1× PBS.
11. Centrifugation is done at 10,000 rpm for 20 min at 4°C.
12. A vacuum is used to remove as much oil as possible and only a small amount of PBS.
13. The volume is completed to 200 mL with 1× PBS, and agar beads are resuspended by manual shaking.
14. This washing step is repeated, and this time half the PBS is removed.
15. After the last washing step, most of the PBS is removed and gives a volume of approximately 10 mL.
16. Beads are resuspended and ready to be injected. Agar beads are conserved at 4°C and can be used up to 1 month.

3.3.2.3. Determination of Colony-Forming Units Prior to Injection

1. An aliquot of 1 mL agar beads is added to 9 mL PBS; this dilution is homogenized with a Polytron for 30 s at maximum speed. The apparatus is sterilized after each sample by a short burst in ethanol 70% and in sterile water.
2. An aliquot of 100 µL is diluted serially to 10^{-4} on a BHIA plate.
3. Plates are incubated overnight at 37°C, and colony-forming units are determined.

It should be noted here that one is targeting an agar-bead preparation containing 10^5 to 10^6 CFUs/100 µL to be injected. To complete the actual screening with 72 different STM mutants in the rat lung for 7 days, a minimum of 10^6 total bacteria is required. Hence, it is critical that all clones are represented at the same level when attempting to produce a chronically infected animal (10^4 minimum per STM mutant × 72 mutants per animal).

3.3.2.4. Inoculation into Animals

Male Sprague-Dawley rats, 450 to 500 g in weight, are used according to the ethics committee for animal treatment. The animals are anesthetized using isoflurane and inoculated by intubation using an 18-G venous catheter and a syringe (1-mL tuberculin) with 120 µL of a suspension of agar beads containing 10^6 colony-forming units (CFU) of bacteria. After 7 days, lungs are removed from sacrificed rats, and homogenized tissues are plated in triplicate on PIA for total number of *P. aeruginosa* bacterial cells and MHA supplemented with antibiotics.

1. After the appropriate *in vivo* incubation time of 7 days, animals are sacrificed, and bacteria are recovered from the targeted organs.

2. Tissues are recovered by dissection and homogenized with a Polytron homogenizer in 10 mL sterile 1× PBS, pH 7.0, contained in a 50-mL falcon tube.
3. 100 µL of homogenized tissues are plated on MHA. After the *in vivo* selection, 10^4 colonies recovered from a single PIA plate are pooled in 5 mL PBS. From the 5 mL, 1 mL is spun down and resuspended in 1 mL of TE PCR (the *in vivo* pool).
4. The *in vivo* pool is boiled 10 min and spun down, and 10 µL of supernatant is used in PCR analysis as described above. Ten microliters of PCR are used for 1% agarose gel electrophoresis separation.
5. PCR amplification products of tags present in the *in vivo* pool are compared with amplified products of tags present in the *in vitro* pool (**Fig. 3B**). We use a multiplex PCR approach combining the different amplified product sizes and confirm negative clones using specific primer sets in single PCR assays (***17***).
6. Mutants that give PCR amplicon from the *in vitro* pool and not from the *in vivo* pools are purified and kept for further analysis.

3.4. Cloning and Analysis of Disrupted Genes from Attenuated Mutants

Instead of using inverse PCR and on a routine basis, chromosomal DNA from attenuated mutants is prepared using the QIAGEN genomic DNA extraction kit as described in the manufacturer's protocol.

1. Chromosomal DNA (1 to 5 µg) is digested with endonuclease (in our case *Pst*I), giving a large range of fragment sizes.
2. Digested chromosomal DNA is cloned into pTZ18R predigested with the corresponding endonuclease and ligation reactions are done as follows:
3. 1 µg of digested chromosomal DNA is mixed with 50 ng of digested pTZ18R in 20 µL of 1× T_4 DNA ligase buffer with 40 units of T_4 DNA ligase.
4. Incubate overnight at 16°C.
5. Ligated products are purified using microcon PCR (Millipore) as described by the manufacturer's instructions and resuspended in 5 µL H_2O.
6. The 5-µL recombinant plasmid solution is used for electroporation in *E. coli* ElectroMAX DH10B as recommended by the manufacturer.
7. After the electroporation, cells are spun down and resuspended in 100 µL of BHI to be plated on a selective plate. Colonies are recovered by scraping from the place using 5 mL BHI.
8. Bacteria containing pTZ18R containing an insertion of genomic DNA encoding the transposon antibiotic resistance marker from mini-Tn5Km or mini-Tn5Tc and mini-Tn5GFP are plated on TSA with Km (50 µg/mL) or Tc (20 µg/mL), respectively.
9. Clones are kept and purified for plasmid analysis.
10. Plasmid DNA is prepared with QIAGEN midi preparation kit as described by the manufacturer.
11. These plasmids are sequenced using the complementary primer of the corresponding tagged mutant or the three conserved transposon primers encoding antibiotic resistance. Automated sequencing is done as suggested by the manufacturer.
12. DNA sequences obtained are assembled and subjected to database searches using BLAST included in the GCG Wisconsin package (version 11.0). Similarity searches with complete genomes can be performed at NCBI using the microbial genome sequences at http://www.ncbi.nlm.nih.gov, or in this specific case for *P. aeruginosa*, http://www.pseudomonas.com.

3.5. Construction of Gene Knockouts for Selected STM Mutants

Because it is well-known that transposon insertions may give polar mutations (except for insertions in genes at the end of an operon), a method is essential to construct gene knockouts in *P. aeruginosa* giving a clean genetic background. However, despite the development of many genetic tools for *P. aeruginosa* over the past decade, isolation of defined deletion mutants is still a relatively tedious process that relies on construction of deletion alleles, most often tagged with an antibiotic-resistance gene, on a suicide plasmid, followed by recombination of the plasmid-borne deletions into the chromosome, usually after conjugal transfer of the suicide plasmid (***24***).

PCR and recombinational technologies can be exploited to substantially accelerate virtually all steps involved in the gene-replacement process. We now use a novel method for rapid generation of unmarked *P. aeruginosa* deletion mutants. The method was applied to deletion of 25 *P. aeruginosa* genes encoding transcriptional regulators of the GntR family (***15***).

The method that we now use can be summarized as follows: Three partially overlapping DNA fragments are amplified and then spliced together *in vitro* by overlap extension PCR. The resulting DNA fragment is cloned *in vitro* into the Gateway vector pDONR221 and then recombined into the Gateway-compatible gene-replacement vector pEX18ApGW. The plasmid-borne deletions are next transferred to the *P. aeruginosa* chromosome by homologous recombination. Unmarked deletion mutants are finally obtained by Flp-mediated excision of the antibiotic resistance marker. The protocol below is essentially as developed by Choi and Schweizer (***15***) and is summarized in **Figure 5** with technical details below. The specific example used is for a deletion on the PA2896 gene isolated by STM with details confirming the PA2896 deletion by PCR in **Figure 6** and analysis in CI in **Figure 7**.

3.5.1. First-Round PCRs for PCR Amplification of the Gm Resistance Gene Cassette

1. A 50-μL PCR reaction contained 5 ng pPS856 template DNA, 1× HiFi Platinum *Taq* buffer, 2 mM $MgSO_4$, 200 μM dNTPs, 0.2 μM of primer Gm-F and Gm-R, and 5 units of HiFi Platinum *Taq* polymerase (Invitrogen). Cycle conditions are 95°C for 2 min, followed by 30 cycles of 94°C for 30 s, 50°C for 30 s, and 68°C for 1 min 30 s, and a final extension at 68°C for 7 min.
2. The resulting 1053-bp PCR product is purified by agarose gel electrophoresis and its concentration determined spectrophotometrically using the Quant-it Picogreen kit (Invitrogen).

3.5.1.1. PCR Amplification of 5′ and 3′ Gene Fragments

Two 50-μL PCR reactions are prepared.

1. The first reaction contains 20 ng chromosomal template DNA, 1× HiFi Platinum *Taq* buffer, 2 mM $MgSO_4$, 5% DMSO, 200 μM dNTPs, 0.8 μM of PA2896-UpF-GWL and PA2896-UpR-Gm primers for the constructed deletion of PA2896, and 5 units of HiFi Platinum Taq polymerase.
2. The second reaction contains the same components as the first except for 0.8 μM of PA2896-DnF-Gm and PA2896-DnR-GWR. Cycle conditions are 94°C for 5 min, followed

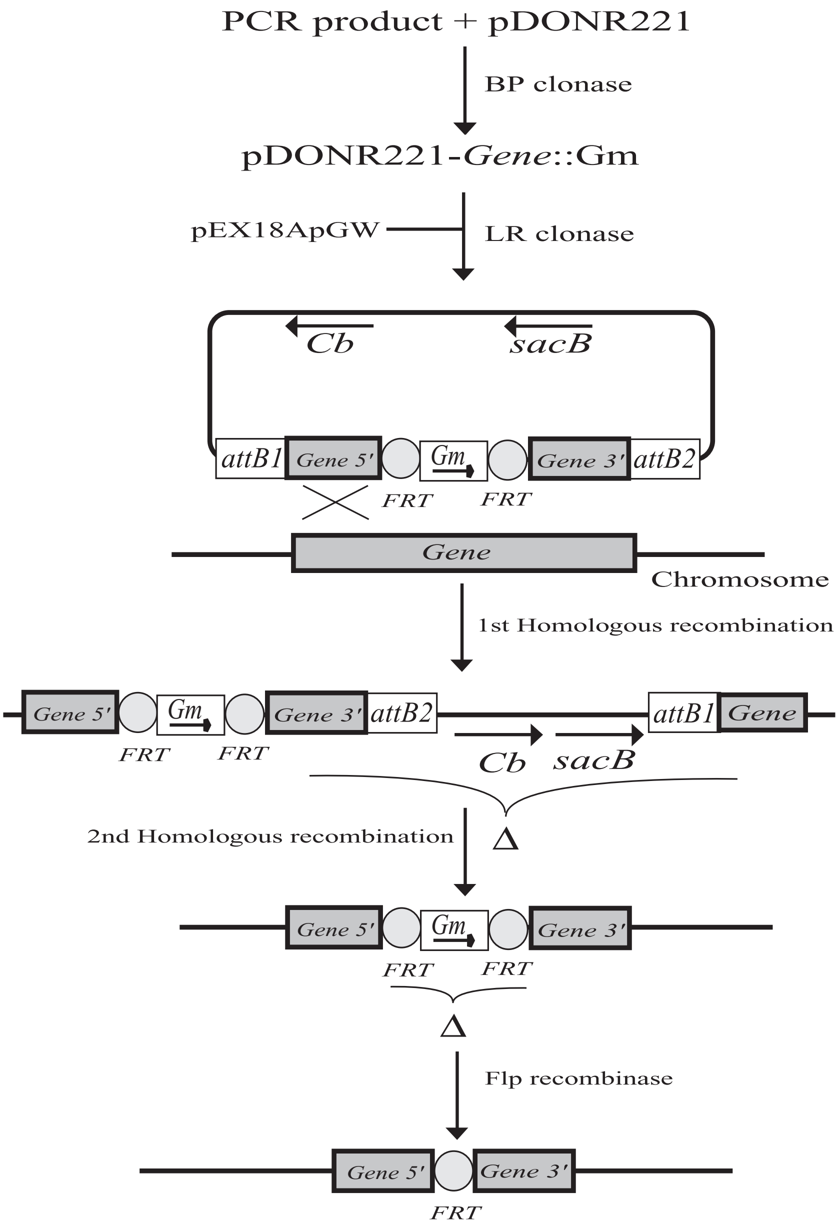


Fig. 5. General scheme for construction of *P. aeruginosa* knockout mutants: Gateway-recombinational cloning and return of the plasmid-borne deletion allele to the *P. aeruginosa* chromosome. The mutant DNA fragment generated by overlap extension PCR is first cloned into pDONR221 via the BP clonase reaction to create the entry clone pDONR221-*Gene*::Gm, which then serves as the substrate for LR clonase–mediated recombination into the destination vector pEX18ApGW. The resulting suicide vector pEX18ApGW-*Gene*::Gm is then transferred to *P. aeruginosa*, and the plasmid-borne deletion mutation is exchanged with the chromosome to generate the desired deletion mutant. Please note that, as discussed in the text, gene replacement by double crossover can occur quite frequently, but it can also be a rare event, in which case allele exchange happens in two steps involving homologous recombination. First, the suicide plasmid is integrated via a single-crossover event resulting in generation of a merodiploid containing the wild-type and mutant allele. Second, the merodiploid state is resolved by *sacB*-mediated sucrose counterselection in the presence of gentamicin, resulting in generation of the illustrated chromosomal deletion mutant. An unmarked mutant is then obtained after Flp recombinase-mediated excision of the Gm marker (*15*).

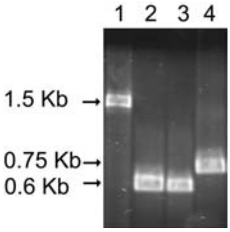


Fig. 6. Allelic replacement analysis by PCR. PCR reactions were done as described in **Section 3.5.4** using primers PA2896-UpF-GWL and PA2896-DnR-GWR. Colony PCR was performed on PAO1Δ*PA2896::FRT-Gm-FRT* (lane 1), PAO1Δ*PA2896::FRT* clone A (lane 2), PAO1Δ*PA2896::FRT* clone B (lane 3), and PAO1 wild type (lane 4). The sizes of the expected PCR DNA fragments are indicated.

by 30 cycles of 94°C for 30 s, 56°C for 30 s, and 68°C for 30 s, and a final extension at 68°C for 10 min.

3. The resulting PCR products are purified by agarose gel electrophoresis using QIAquick gel extraction kit and their concentrations determined spectrophotometrically.

3.5.2. Second-Round PCR

1. A 50-µL PCR reaction contains 50 ng each of the PA2896 in 5′ and 3′ purified template DNA and 50 ng of *FRT*-Gm-*FRT* template DNA prepared during first-round PCR. The

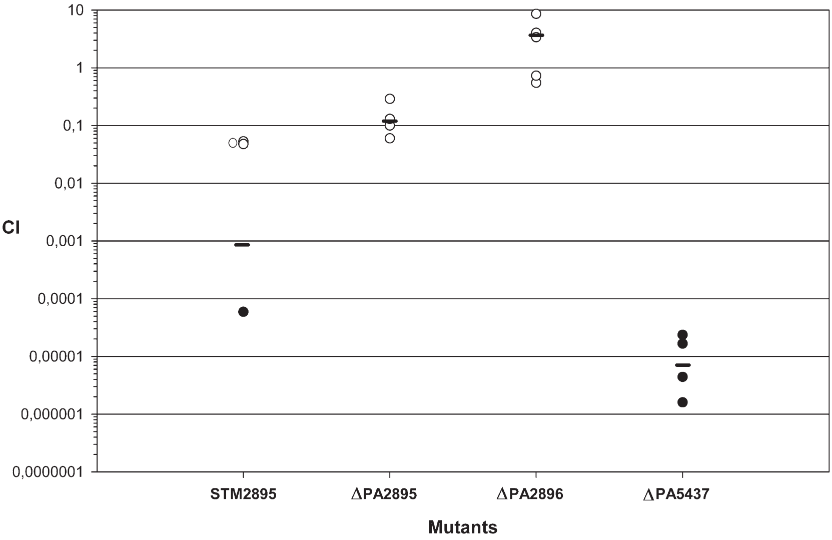


Fig. 7. Competitive index (CI) analysis of *P. aeruginosa* STM and knockout mutants obtained in the rat lung model of chronic infection. The *in vivo* CIs are calculated as previously described (*2, 26*). Each circle represents the CI for a single rat in each set of competitions. A CI of less than 1 indicates a virulence defect. Dark circles indicate that no mutant bacteria were recovered from that animal, and 1 was substituted in the numerator when calculating the CI value. The geometric mean of the CI for all rats in a set of competitions is shown as a solid line. The *in vivo* competitive results for each of the tested strains are as follows: STM2895, 0.0092; ΔPA2895, 0.12; ΔPA2896, 2.84; ΔPA5437 (PycR), 0.0000073 (*8*).

reaction mix also contains 1× HiFi Platinum *Taq* buffer, 2 mM $MgSO_4$, 5% DMSO, 200 µM dNTPs, and 5 units of HiFi Platinum *Taq* polymerase. After an initial denaturation at 94°C for 2 min, 3 cycles of 94°C for 30 s, 55°C for 30 s, and 68°C for 1 min are run without added primers. The third cycle is paused at 30 s of the 68°C extension, primers GW-*attB1* and GW-*attB2* are added to 0.2 µM each, and the cycle is then finished by another 30-s extension at 68°C. The PCR is completed by 25 cycles of 94°C for 30 s, 56°C for 30 s, and 68°C for 5 min and a final extension at 68°C for 10 min.

2. The resulting major PCR product is purified by agarose gel electrophoresis and its concentrations determined spectrophotometrically. The identity of the PCR fragment is confirmed by *Xba*I digestion (each *FRT* site of the *FRT*-Gm-*FRT* fragment contains an *Xba*I site).

3.5.3. BP and LR Clonase Reactions

1. The BP and LR clonase reactions for recombinational transfer of the PCR product into pDONR221 and pEX18ApGW, respectively, are performed as described in Invitrogen's Gateway cloning manual, but using only half of the recommended amounts of BP and LR clonase mixes and *E. coli* One Shot MAX Efficiency DH5α-T1^r.
2. The presence of the correct fragments in transformants obtained with DNA from either clonase reaction was verified by digestion with *Xba*I because each *FRT* site flanking the Gm^r gene contains an *Xba*I site.
3. However, before plasmid isolation from transformants obtained with DNA from the LR clonase reaction, 25 to 50 transformants were (a) patched on LB+Km and LB+Ap plates and (b) simultaneously purified for single colonies on LB+Ap plates. This was necessary to distinguish between those colonies containing only the desired pEX18ApGW-*Gene*::Gm from those containing this plasmid and the frequently contaminating pDONR-*Gene*::Gm (pEX18Ap-derived plasmids confer Ap^r, and pDONR plasmids confer Km^r).

3.5.4. Transfer of Plasmid-Borne Deletions to the P. aeruginosa *Chromosome*

An electroporation method is used to transfer the pEX18ApGW-borne deletion mutations to *P. aeruginosa.*

1. Briefly, 6 mL of an overnight culture grown in LB medium was harvested in four microcentrifuge tubes by centrifugation (1 to 2 min, 16,000 ×*g*) at room temperature.
2. Each cell pellet was washed twice with 1 mL of room-temperature 300 mM sucrose, and they were then combined in a total of 100 µL 300 mM sucrose.
3. For electroporation, 300 to 500 ng of plasmid DNA was mixed with 100 µL of electrocompetent cells and transferred to a 2-mm-gap-width electroporation cuvette. After applying a pulse (settings: 25 µF; 200 ohm; 2.5 kV on a Bio-Rad GenePulser), 1 mL of LB medium was added at once, and the cells were transferred to a polystyrene tube and incubated for 1 h at 37°C.
4. The cells were then harvested in a microcentrifuge tube. Eight hundred microliters of the supernatant was discarded and the cell pellet resuspended in the residual medium.
5. The entire mixture was then plated on two LB plates containing 30 µg per mL Gm (LB+Gm30). The plates were incubated at 37°C until colonies appeared (usually within 24 h). Under these conditions, the transformation efficiencies were generally 30 to 100 transformants per µg of DNA.
6. A few colonies were patched on LB+Gm30 plates and LB+Cb200 plates to differentiate single- from double-crossover events.

7. To ascertain resolution of merodiploids, Gm^r colonies were struck for single colonies on LB+Gm30 plates containing 5% sucrose. Gm^r colonies from the LB-Gm-sucrose plates were patched onto LB+Gm30+5% sucrose, as well as LB plates with 200 µg/mL carbenicillin (LB+Cb200). Colonies growing on the LB-Gm-sucrose but not on the LB-carbenicillin plates were considered putative deletion mutants.
8. The presence of the correct mutations was verified by colony PCR. To do this, a single large colony (or the equivalent from a cell patch) was picked from an LB-Gm-sucrose plate, transferred to 100 µL TE PCR in a microcentrifuge tube, and boiled for 10 min.
9. Cell debris was removed by centrifugation in a microcentrifuge (2 min; 13,000 × *g*), and the supernatant was transferred to a fresh tube, which was placed on ice.
10. Ten microliters of the supernatant was used as source of template DNA in a 50-µL PCR reaction containing *Taq* buffer, 1.5 mM $MgSO_4$, 5% DMSO, 0.6 µM each of the 5′ and 3′ primers (PA2896-UpF-GWL and PA2896-DnR-GWR), 200 µM dNTPs, and 5 units Hot-StartTaq DNA polymerase. Cycle conditions were 95°C for 15 min, followed by 30 cycles of 95°C for 45 s, 55°C for 30 s, and 72°C for 2 min and a final extension at 72°C for 10 min. PCR products were analyzed by agarose gel electrophoresis.

3.5.5. Flp-Mediated Marker Excision

1. Electrocompetent cells of the newly constructed mutant strain were prepared as described in the preceding paragraph and transformed with 20 ng pFLP2 DNA as described above.
2. After phenotypic expression at 37°C for 1 h, the cell suspension was diluted 1 : 1000 and 1 : 10,000 with either LB or 0.9% NaCl, and 50 µL aliquots were plated on LB+Cb200 plates and incubated at 37°C until colonies appeared.
3. Transformants were purified for single colonies on LB+Cb200 plates. Ten single colonies were tested for antibiotic susceptibility on LB ± Gm30 plates and on an LB+Cb200 plate.
4. Two Gm^s Cb^r isolates were struck for single colonies onto an LB+5% sucrose plate and incubated at 37°C until sucrose-resistant colonies appeared. Ten sucrose-resistant colonies were retested on an LB+5% sucrose (master) plate and an LB+Cb200 plate.
5. Finally, two sucrose-resistant and Cb^s colonies were struck on LB plates without antibiotics and their Cb^s and Gm^s phenotypes confirmed by patching on LB ± Cb200 and LB ± Gm30 plates.
6. Deletion of the Gm^r marker was assessed by colony PCR utilizing the conditions and primers described above.

3.6. Competitive Index Analysis

The competitive index (CI) is a sensitive measure of the relative degree of virulence attenuation of a particular mutant in mixed infection with the wild-type strain. It is defined as the ratio of the mutant strain to the wild type in the output divided by the ratio of the two strains in the input (*25, 26*). In addition to these studies, it is crucial to determine the *in vitro* growth curve of knockouts along with the wild type (*in vitro* CI) to confirm that the clones isolated have no bias by having mutations in genes affecting generation time and growth and in being out-competed by the wild type. Growth curves from each *P. aeruginosa* knockout mutant are constructed at 1-h time points for a period of 18 h in TSB broth using serial dilutions of colony-forming units; clones retained should have the same growth pattern as the wild type. This is the case for the STM and knockout mutants presented in **Figure 7** and prior to estimating the CI analysis *in vivo*.

Also, knockout mutants selected will be screened by auxanography on minimal media to eliminate attenuated strains having growth defects.

It is only after this initial screening that one may estimate the relative pathogenicity of selected knockout mutants constructed by determination of the competitive infectivity index test.

Bacterial cells embedded in agarose beads were prepared as described (*8, 23*), and the scheme is presented in **Figure 4**. Male Sprague-Dawley rats of approximately 500 g in weight are used according to the recommendations of the ethics committee for animal treatment. The animals are anesthetized using isoflurane; inoculation into the lungs is done by intubation using an 18-G venous catheter and a syringe (1-mL tuberculin) containing 120 µL of an agarose bead suspension with a total of 10^6 bacterial cells. Seven days after infection, animals are sacrificed; their lungs are removed, and homogenized tissues are plated on PIA and MHA agar. The wild-type strain is differentiated from STM or knockout mutants using Cb resistance encoded by the pUCP19 plasmid (**Note 4**).

1. The wild-type strain colony forming units (CFU) are determined on MHA plates containing Cb. PIA is used to determine total bacterial counts.
2. A colony from a fresh plate is used to inoculate 50 mL TSB in a 250-mL Erlenmeyer flask. A culture from the wild-type strain PAO1 containing the pUCP19 plasmid and a culture from each mutant strain are grown overnight at 37°C with agitation at 250 rpm. Bacterial growth is monitored at an OD_{600} until 1.0 is obtained and which yields 2×10^{10} CFU/mL.
3. A 200-µL aliquot of the overnight culture is completed to 1 mL with fresh TSB in a 1.5-mL microtube to give a final concentration of approximately ~1×10^{10} CFU/mL.
4. A 250-µL aliquot of the wild-type strain dilution is mixed with 250 µL of a mutant strain dilution and added to 4.5 mL of TSB in a 15 × 150 mm culture tube.
5. The 5-mL aliquot is mixed in a 50-mL tube containing 20 mL of 2% sterile agarose (Nusieve GTG; FMC) in 1× PBS at 48°C.
6. The agarose-broth mixture is added to a 250-mL Erlenmeyer flask containing 200 mL of heavy mineral oil at 48°C and rapidly stirred on a magnetic stirrer in a water bath (setting 500 to 600 rpm) on a hotplate stirrer (model M13; Staufen) as depicted in **Figure 4**.
7. The mixture is cooled gradually with ice chips to 0°C in a period of 5 min. The agarose beads are transferred into a sterile 500-mL Squibb-type separator funnel and washed once with 200 mL 0.5% deoxycholic acid sodium salt (SDC) in PBS, once with 200 mL 0.25% SDC in PBS, and three times with 200 mL PBS. The bead slurry is allowed to settle, and a 50-mL sample was recovered.
8. For the final wash, a minimal volume of approximately 20 mL of bead slurry is recovered. Agarose beads are incubated in a 50-mL tube on ice, and the remaining PBS is removed so as to concentrate beads to a final volume of approximately 15 mL.
9. Sterile agarose beads are stored at 4°C and can be used for several experiments; bacterial counts are maintained up to 1 month.
10. One milliliter of bead slurry is diluted in 9 mL PBS and homogenized (Polytron), and serial dilutions are plated on PIA and on MHA supplemented with Cb or Gm. Colony-forming units are determined after 18 h at 37°C and are used to calculate the input ratio of mutant to wild-type bacterial cells.
11. After the *in vivo* passage, colony-forming units on plates represent the total number of bacteria present in the rat lungs. Colonies that grew on MHA+Cb represent the number of

wild-type PAO1 bacteria. Colonies obtained on MHA+Gm represent the number of mutant bacteria. Colonies on PIA represent the total number of *P. aeruginosa* bacterial cells in the rat lung.

12. The CI is defined as the CFU output ratio of mutant when compared with wild-type strain, divided by the CFU input ratio of mutant to wild-type strain (*25, 26*). The final CI is calculated as the geometric mean for animals in the same group, and experiments are done at least in triplicate (*26*). Each *in vivo* competition is tested for statistical significance by Student's two-tailed *t*-test (*26*).

The examples that we use here are the STM2895, ΔPA2895, ΔPA2896, and ΔPA5437 (bcxR) (*8*) for analysis of CI values. As depicted in **Figure 7**, the STM2895 and ΔPA2895 have CI values of 0.001 and 0.1, whereas the ΔPA2896 has a CI value of 2. In contrast, the ΔPA5437 has a CI of 0.00007 when compared with the wild type.

Notes

1. It might be necessary to screen several colonies to find a correct recombinant. It is possible to pool several colonies to reduce the number of PCRs (*17*). To bypass the necessity of doing plasmid preparations, PCR can be done on bacterial cell lysates. One or several colonies are resuspended in 100 μL TE PCR buffer, boiled 10 min, and spun down. Ten microliters of supernatant are used as PCR template.
2. In a defined library, each mutant has the same tag but is assumed to be inserted at a different location in the bacterial chromosome. Prior to starting STM, Southern blot hybridization is necessary to confirm the random integration of the mini-Tn*5*.
3. Parameters concerning each different animal model should be well defined. The inoculum size necessary to cause infection determines the complexity of mutants to be pooled. In fact, each mutant in a defined input pool has to be in a sufficient cell number to initiate infection. The inoculum size must not be too high, resulting in the growth of mutants that would otherwise have not been detected. Other important parameters in STM include the route of inoculation and the time course of a particular infection. Also, certain gene products important directly or indirectly for initiation or maintenance of the infection may be niche-dependent or expressed specifically in certain tissues only. If the duration of the infection is short, genes important for establishment of the infection will be found, and if the duration is long, genes important for maintenance of infection will be identified.
4. Each STM attenuated mutant has to be confirmed by a second round of STM screening, comparisons between *in vivo* bacterial growth rate of mutants versus growth of the wild type in single or competitive infections, or estimation of LD_{50}.

Acknowledgments

The authors thank H.P. Schweizer for strains, plasmids, and support in using the gene knockout system, D. Woods for the suggestions and comments for the rat model of chronic lung infection, and R.W. Hancock for constant encouragement and support. This work was supported by a Canadian Institute for Health Research operating grant in genomics and by a scholar of exceptional merit salary support from le Fonds de Recherche en Santé du Québec to R.C. Levesque and a FRSQ studentship to I. Kukavica-Ibrulj.

References

1. Handfield, M., and Levesque, R. C. (1999) Strategies for isolation of in vivo expressed genes from bacteria. *FEMS Microbiol. Rev.* **23**, 69–91.
2. Levesque, R. C. (2006) *In Vivo Functional Genomics of Pseudomonas: PCR-Based Signature-Tagged Mutagenesis*. Boston: Springer, pp. 99–120.
3. Handfield, M., Lehoux, D. E., Sanschagrin, F., Mahan, M. J., Woods, D. E., and Levesque, R. C. (2000) *In vivo*-induced genes in *Pseudomonas aeruginosa*. *Infect. Immun.* **68**, 2359–2362.
4. Hensel, M., Shea, J. E., Gleeson, C., Jones, M. D., Dalton, E., and Holden, D. W. (1995) Simultaneous identification of bacterial virulence genes by negative selection. *Science* **269**, 400–403.
5. Lehoux, D. E., Sanschagrin, F., and Levesque, R. C. (1999) Defined oligonucleotide tag pools and PCR screening in signature-tagged mutagenesis of essential genes from bacteria. *Biotechniques* **26**, 473–478, 480.
6. Autret, N., and Charbit, A. (2005) Lessons from signature-tagged mutagenesis on the infectious mechanisms of pathogenic bacteria. *FEMS Microbiol. Rev.* **29**, 703–717.
7. Potvin, E., Lehoux, D. E., Kukavica-Ibrulj, I., Richard, K. L., Sanschagrin, F., Lau, G. W., and Levesque, R. C. (2003) *In vivo* functional genomics of *Pseudomonas aeruginosa* for high-throughput screening of new virulence factors and antibacterial targets. *Environ. Microbiol.* **5**, 1294–1308.
8. Kukavica-Ibrulj, I., Peterson, A., Sanschagrin, F., Whiteley, M., and Levesque, R. C. (2006) Functional genomics of PycR, a major LysR family transcriptional regulator essential for maintenance of *Pseudomonas aeruginosa* in the lung. *Microbiology*, accepted, in revision.
9. Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., et al. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406**, 959–964.
10. Winsor, G. L., Lo, R., Sui, S. J., Ung, K. S., Huang, S., Cheng, D., et al. (2005) *Pseudomonas aeruginosa* Genome Database and PseudoCAP: facilitating community-based, continually updated, genome annotation. *Nucleic Acids Res.* **33**, D338–343.
11. Wiehlmann, L., Salunkhe, P., Larbig, K., Ritzka, R., and Tummler, B. (2002) Signature-tagged mutagenesis of *Pseudomonas aeruginosa*. *Genome Lett.* **1**, 131–139.
12. Lehoux, D. E., Sanschagrin, F., and Levesque, R. C. (2001) Discovering essential and infection-related genes. *Curr. Opin. Microbiol.* **4**, 515–519.
13. De Lorenzo, V., Herrero, M., Jakubzik , U., and Timmis, K. N. (1990) Mini-Tn5 transposon derivatives for insertion mutagenesis, promoter probing, and chromosomal insertion of cloned DNA in Gram-negative eubacteria. *J. Bacteriol.* **172**, 6568–6572.
14. Matthysse, A. G., Stretton, S., Dandie, C., McClure, N. C., and Goodman, A. E. (1996) Construction of GFP vectors for use in gram-negative bacteria other than *Escherichia coli*. *FEMS Microbiol. Lett.* **145**, 87–94.
15. Choi, K. H., and Schweizer, H. P. (2005) An improved method for rapid generation of unmarked *Pseudomonas aeruginosa* deletion mutants. *BMC Microbiol.* **5**, 30.
16. Schweizer, H. P. (1991) Improved broad-host-range lac-based plasmid vectors for the isolation and characterization of protein fusions in *Pseudomonas aeruginosa*. *Gene* **103**, 87–92.
17. Dewar, K., Sabbagh, L., Cardinal, G., Veilleux, F., Sanschagrin, F., Birren, B., and Levesque, R. C. (1998) *Pseudomonas aeruginosa* PAO1 bacterial artificial chromosomes: strategies for mapping, screening, and sequencing 100 kb loci of the 5.9 Mb genome. *Microb. Comp. Genomics* **3**, 105–117.

18. Kwok, S., Kellogg, D. E., McKinney, N., Spasic, D., Goda, L., Levenson, C., and Sninsky, J. J. (1990) Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. *Nucleic Acids Res.* **18**, 999–1005.
19. Rychlik, W. (1993) *Selection of primers for polymerase chain reaction*. Totowa, NJ: Humana Press.
20. Sambrook, J., and Russell, D. W. (2001) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
21. Simon, R., Priefer, U., and Pühler, A. (1983) A broad host range mobilization system for *in vivo* genetic engineering: transposon mutagenesis in gram negative bacteria. *Bio/Technology* **1**, 784–791.
22. Cash, H. A., Woods, D. E., McCullough, B., Johanson, W. G. Jr., and Bass, J. A. (1979) A rat model of chronic respiratory infection with *Pseudomonas aeruginosa*. *Am. Rev. Respir. Dis.* **119**, 453–459.
23. van Heeckeren, A. M., and Schluchter, M. D. (2002) Murine models of chronic *Pseudomonas aeruginosa* lung infection. *Lab. Anim.* **36**, 291–312.
24. Schweizer, H. P. (1992) Allelic exchange in *Pseudomonas aeruginosa* using novel ColE1-type vectors and a family of cassettes containing a portable oriT and the counter-selectable *Bacillus subtilis* sacB marker. *Mol. Microbiol.* **6**, 1195–1204.
25. Beuzon, C. R., and Holden, D. W. (2001) Use of mixed infections with *Salmonella* strains to study virulence genes and their interactions in vivo. *Microbes Infect.* **3**, 1345–1352.
26. Hava, D. L., and Camilli, A. (2002) Large-scale identification of serotype 4 *Streptococcus pneumoniae* virulence factors. *Mol. Microbiol.* **45**, 1389–1406.

6

Whole-Genome Detection of Conditionally Essential and Dispensable Genes in *Escherichia coli* via Genetic Footprinting

Michael D. Scholle and Svetlana Y. Gerdes

Summary

We present a whole-genome approach to genetic footprinting in *Escherichia coli* using Tn*5*-based transposons to determine gene essentiality. A population of cells is mutagenized and subjected to outgrowth under selective conditions. Transposon insertions in the surviving mutants are detected using nested polymerase chain reaction (PCR), agarose gel electrophoresis, and software-assisted PCR product size determination. Genomic addresses of these inserts are then mapped onto the *E. coli* genome sequence based on the PCR product lengths and the addresses of the corresponding genome-specific primers. Gene essentiality conclusions were drawn based on a semiautomatic analysis of the number and relative positions of inserts retained within each gene after selective outgrowth.

Key Words: dispensable genes; *E. coli*; essential genes; genetic footprinting; genome; Tn*5*; transposome; transposon mapping; transposon mutagenesis.

1. Introduction

The transposon-based approach termed *genetic footprinting* was originally developed for the identification of genes essential for viability of *Saccharomyces cerevisiae* under various growth conditions (*1, 2*). The first step in genetic footprinting involves random transposon mutagenesis of a large number of cells to generate a comprehensive population of insertion mutants. This population must be complex enough to include several unique mutations per gene in the genome. The second step is competitive outgrowth of the mutagenized population under relevant selective conditions. The final step includes analysis of individual mutants surviving in the population using direct sequencing across insertion junctions or various PCR- or hybridization-based techniques. The loss of mutants after selective outgrowth is indicative of the essentiality of the corresponding gene products under experimental growth conditions.

Various modifications of genetic footprinting have been recently applied in several microorganisms: *Mycoplasma genitalium* and *Mycoplasma pneumoniae* (*3*),

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

Haemophilus influenzae (*4, 5*), *Mycobacterium tuberculosis* (*6–8*), *Pseudomonas aeruginosa* (*9, 10*), *Helicobacter pylori* (*11*), *Salmonella typhimurium* (*12*), including several studies in *Escherichia coli* (*13–16*). Genetic footprinting experiments reported in *E. coli* have utilized mini-Tn*10* mutagenesis *in vivo* (*13, 15*) and *in vitro* mutagenesis with Tn*5* delivered via transposomes (*14, 16, 17*).

An explosion in the development of *in vitro* transposition techniques has occurred within the past decade, thus helping circumvent many limitations of the classic *in vivo* approaches. These techniques have also extended the application of transposition tools to previously genetically intractable microorganisms (for reviews, see **Chapter 2** and Refs. *18* and *19*). *In vitro* transposition systems have been developed based on bacteriophage Mu (*20*); bacterial transposons Tn*3* (*21, 22*), Tn*5* (*23, 24*), Tn*7* (*25*), Tn*10* (*26*), and Tn*552* (*27*); yeast transposon Ty*1* (*28*); and *mariner* transposon of insects (*29*). Several of these are available commercially as specialized kits for numerous applications (**Note 1**).

The use of the *in vitro* transposome–based strategy of Tn*5* transposition for analysis of microbial gene essentiality has several advantages over classic *in vivo* transposon mutagenesis: (1) only single irreversible insertions are produced because the only source of transposase activity is within the transposome complex formed *in vitro*; (2) there is no requisite to assemble an elaborate transposon delivery system with tight regulation of replication and transposase expression; and (3) a limit of one insertion per cell can be achieved based on the ratio of transposome complexes to competent cells at the time of transformation.

An important aspect of these transposon-based mutagenesis techniques is electroporation with preformed transposomes using precleaved transposon DNA in a stable complex with a modified transposase. This technology was first developed for Tn*5* (*30, 31*) and is now also available for bacteriophage Mu (*32*). It is based on the discovery of Tn*5* synaptic complex formation whereby cleavage of the transposon DNA by transposase can be separated in time from the actual transposition event (*31*). This precleaved transposon DNA forms a stable synaptic complex (transposome) in the absence of divalent metal ions. Transposomes can then be electroporated into electrocompetent target cells where they “jump” into genomic (or extrachromosomal) DNA in the presence of intracellular Mg^{2+} (*24*).

This chapter describes the protocol for an experimental detection of a nearly complete list of the *E. coli* genes essential and dispensable under specific environmental and genetic conditions via a transposome-based genetic footprinting technique (*14, 17*). Aerobic logarithmic growth in complex rich medium is used as an example.

2. Materials

1. pMOD EZ::TN<Kan2> (Epicentre Technologies, Madison, WI) or other Tn*5* transposon.
2. Tn*5* hyperactive transposase (Epicentre).
3. *Pvu* II restriction nuclease.
4. QiaQuick Gel Extraction Kit (Qiagen, Valencia, CA).
5. Transposome formation buffer: 40 mM Tris-acetate (pH 7.5), 100 mM potassium glutamate, 0.1 mM EDTA, 1 mM dithiothreitol, and tRNA (0.1 mg/mL).
6. 0.025-µM dialysis filters (Millipore, Bedford, MA).

7. Microcon Centrifugal Filter Device YM-100 (Millipore).
8. Enriched Luria-Bertani (LB) medium composed of 10 g tryptone/liter, 5 g yeast extract/liter, 50 mM NaCl, 9.5 mM NH_4Cl, 0.528 mM $MgCl_2$, 0.276 mM K_2SO_4, 0.01 mM $FeSO_4$, 5×10^{-4} mM $CaCl_2$, and 1.32 mM K_2HPO_4. The growth medium also included the following micronutrients: 3×10^{-6} mM $(NH_4)_6(MoO_7)_{24}$, 4×10^{-4} mM H_3BO_3, 3×10^{-5} mM $CoCl_2$, 10^{-5} mM $CuSO_4$, 8×10^{-5} mM $MnCl_2$, and 10^{-5} mM $ZnSO_4$ (final concentrations).
9. SOB: 20 g peptone, 5 g yeast extract, 0.584 g NaCl, and 0.186 g KCl in 1 L of distilled H_2O (autoclaved).
10. SOC: Autoclaved SOB plus the following components added separately from filter-sterilized stock solutions: 10 mM $MgCl_2$, 10 mM $MgSO_4$, and 20 mM glucose (final concentrations).
11. 15% glycerol (2 L; autoclaved).
12. Kanamycin at the final concentration of 10 μg/mL.
13. Environmental shaker or fermentor.
14. *E. coli* genomic DNA isolation kit available from Fermentas (Nahover, MD), Bio-Rad (Hercules, CA), BD Biosciences-Clontech (Palo Alto, CA), Qiagen (Valencia, CA), or Sigma (St. Louis, MO).
15. Transposon-specific primers (**Section 3.4.2**).
16. A set of genome-specific primers (**Section 3.4.2**).
17. Primer design software: PrimerSelect (DNASTAR, Inc., Madison, WI), Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3.cgi), or NetPrimer (http://alces.med.umn.edu/websub.html).
18. Advantage cDNA polymerase (Takara Bio-Clontech, Mountain View, CA).
19. Advantage cDNA polymerase buffer (Takara Bio-Clontech): 40 mM Tricine-KOH, pH 9.2, 15 mM potassium acetate, 3.5 mM magnesium acetate, 3.75 μg/mL bovine serum albumin.
20. 10 mM deoxynucleotide triphosphates (dNTPs) mixture (10 mM each of dATP, dCTP, dGTP, and dTTP in 10 mM Tricine-KOH, pH 7.6).
21. Thermocycler.
22. 0.65% agarose gel in Tris-acetate-EDTA (TAE) running buffer (40 mM Tris-acetate, 1 mM EDTA, pH 8.3).
23. 1 kb Plus DNA Ladder (Invitrogen, Carlsbad, CA).
24. DNA agarose gel imaging system and analysis software from Kodak 1D Image Analysis Software (Eastman Kodak, Rochester, NY) or Labworks Software (UVP Inc., Upland, CA).

3. Methods

The methods below outline (1) strain selection, (2) transposome formation and electroporation, (3) outgrowth of the mutagenized *E. coli* culture, (4) genetic footprinting, (5) PCR product size determination, (6) PCR optimization, and (7) whole-genome mapping and gene essentiality determination.

3.1. Selecting a Strain for a Whole-Genome Transposon Mutagenesis Experiment

The development of the transposome-based and other transposon mutagenesis techniques has increased the number of microbial species and strains where gene essentiality studies (genome-wide or local) can now be conducted. There are only two requirements to consider when choosing a microbial strain for the whole-genome gene essentiality studies: the availability of a complete genome sequence and high efficiency of electroporation. Genomic sequence data are essential for mapping transposon insertions and generating genetic footprints. A complete genome sequence, on a single contig, makes the whole-genome essentiality analysis easier, although incomplete sequence data can be used as well with appropriate modifications of a mapping software. Up-to-date lists of the *E. coli* strains with complete or nearly complete genome sequences are maintained by the Enteropathogen Resource Integration Center (ERIC; http://www.ericbrc.org/portal/eric/enteropathogens?id=enteropathogens); Genomes OnLine Database (GOLD; http://www.genomesonline.org/); and the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/Sites/entrez?db=genome).

An efficient electroporation procedure is crucial for generating a large-enough mutant population required to saturate the entire genome with inserts. The actual number of independent mutants necessary to achieve this goal depends on (1) the genome size of the microorganism under study, (2) randomness in target choice of a specific transposase utilized (**Section 3.2**), and (3) variations in susceptibility of different genomic loci to transposition (**Chapter 22**). In practical terms, the actual number of independent insertion mutants required to achieve the desired transposition density can be 5- to 10-fold higher (when using hyperactive Tn*5* transposase) than that estimated from the genome size alone (**Note 2**).

Electroporation efficiencies of the wild-type *E. coli* and recent clinical isolates are generally much lower than that of laboratory strains. Furthermore, transformation with transposome nucleoprotein complexes is approximately 10,000-fold less efficient than electroporation with supercoiled plasmid DNA of similar size. Using the procedure outlined below, we obtained efficiency of transposome electroporation of 5×10^4 transformants per 1 μg of transposon DNA in MG1655 (this number being approximately 50-fold lower than that for the highly transformable lab strain DH10B). Large volumes of competent cells and transposome reaction mixture are needed to overcome these limits in electroporation efficiency. Optimization of electroporation efficiency for each strain of choice is necessary for genome-scale footprinting.

3.1.1. Generating E. coli Electrocompetent Cells

It is important that all labware for competent cell preparation is free of detergents. This can be achieved by rinsing culture flasks with distilled water on a rotary shaker for ~15 min or by autoclaving all glassware with water to remove residual detergents. Remove the distilled H_2O and use the flasks for solution preparation or cell growth.

1. Inoculate a single colony from a fresh LB plate into a 10-mL culture of SOB broth; incubate overnight at 37°C.
2. Meanwhile, 1 L of 15% glycerol should be prepared, autoclaved, and stored at 4°C.
3. Add 10 mL overnight culture to 1 L (1 : 100 dilution) of prewarmed (37°C) SOB broth.
4. Incubate with shaking at 37°C for 2 to 3 h to mid-log growth (OD_{600} of 0.6 to 0.7).
5. Chill flask on wet ice for 10 min. Important: Keep cells on ice from this point onward.
6. Precool the centrifuge and the rotor to 0°C to 4°C.

7. Harvest cells by centrifugation at 6000×*g* at 4°C.
8. Wash pellets twice with 15% glycerol (to remove salts). Be sure to completely resuspend the cell pellet during each wash.
9. Upon the final centrifugation, resuspend the cells in 5 mL (or less) of 15% glycerol.
10. Aliquot, flash freeze, and store at −80°C until needed.

The efficiency of electroporation for *E. coli* strains MG1655 and DH10B should be no less than 10^8 and 5×10^9 colony-forming units (CFU) per 1 µg of pUC18 double-stranded, supercoiled DNA, respectively.

3.2. Transposomes

3.2.1. Commercially Preformed Transposomes

A variety of preformed transposome complexes using hyperactive Tn*5* transposase (**Note 3**) and transposons with different selection markers can be purchased from Epicentre Technologies. However, for a genome-scale transposition experiment in a strain with low electroporation rates (when large transposome quantities are required), the purchase of commercial transposomes can be costly. We recommend the following procedure for generation of transposomes *in vitro* using transposon DNA and hyperactive Tn*5* transposase available in bulk from Epicentre.

3.2.2. Design of an Artificial Transposon for Genetic Footprinting

Practically any DNA fragment over ~200 bp in size flanked by the 19 bp mosaic end (ME) sequences (**Fig. 1**) can be used to form a transposome ***(23, 33)***. Plasmids containing a variety of artificial transposons can be purchased or created in the lab using a transposon construction vector (e.g., pMOD<MCS>; **Fig. 1B, C**). Options for custom transposon-containing plasmids include an optimal selectable marker conferring antibiotic resistance at a single copy per genome, primer annealing sites, or promoter sequences (**Note 4**), which can be inserted between the two ME sequences. Two *Pvu* II restriction sites flank the ME sequences in pMOD<MCS> vectors to allow for precise release of the transposon DNA: **CAG^<u>CTG</u>**<u>TCTCTTATACACATCT</u> (*Pvu* II site is in bold, ME sequence is underlined).

3.2.3. Preparation of Transposon DNA

Precleaved transposon DNA can be generated by restriction enzyme digestion of an appropriate plasmid or by PCR amplification. In the former case, it might be necessary to passage the plasmid through the *E. coli* strain to be mutagenized to avoid restriction-modification incompatibility during transposition. In either case, it is very important that generated transposon DNA molecules have perfectly blunt ends with no additional bases past the MEs. For this reason, the use of PCR primers annealing outside the ME sequences with subsequent *Pvu*II digestion of PCR products is recommended (**Note 5**). Preparation of the EZ::TN<KAN-2> transposon DNA (Epicentre) is described below as an example.

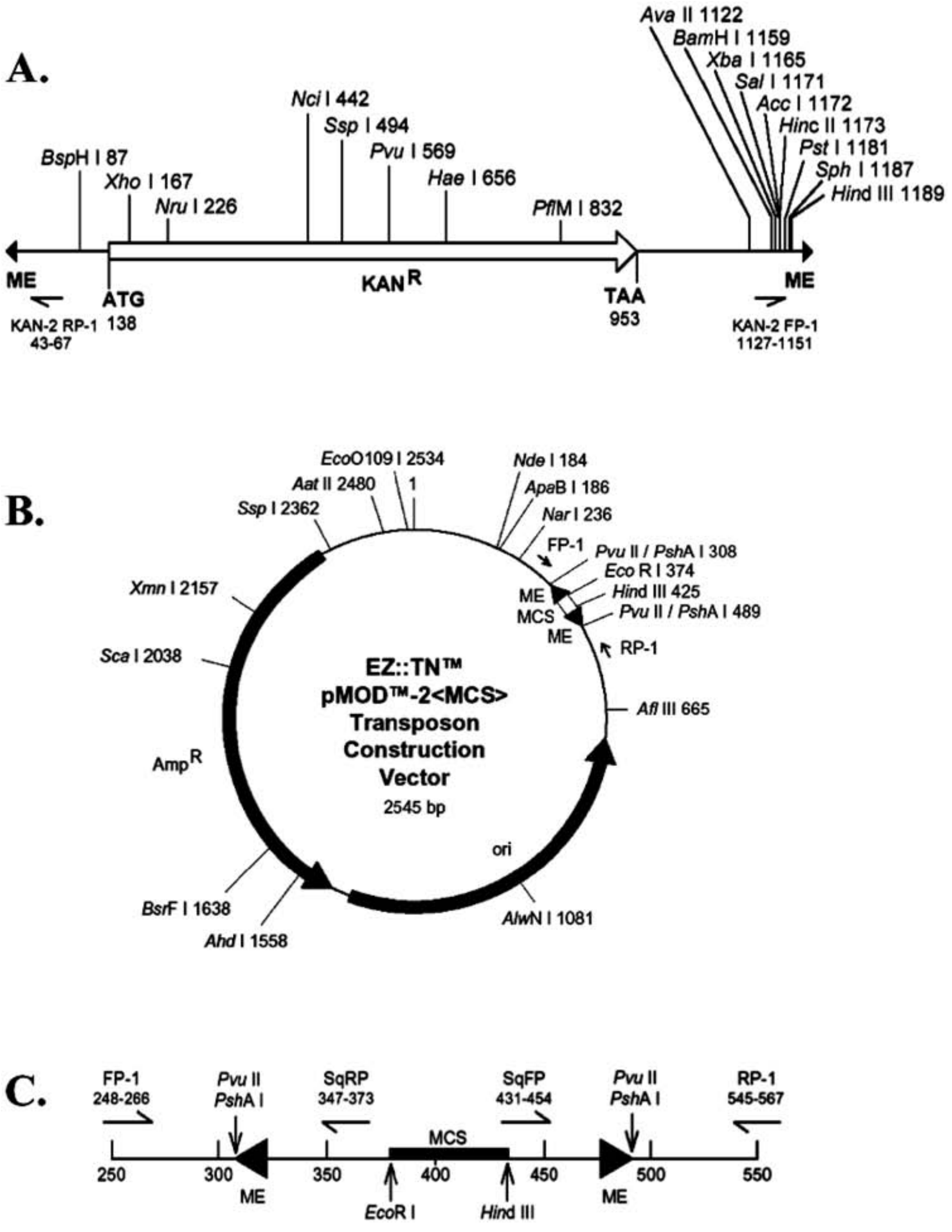


Fig. 1. Maps of Tn5 transposon–related DNA and plasmid. **(A)** Map of Tn5<KAN>^{TM} DNA flanked by mosaic ends (ME) for transposase *Tn*5 attachment. **(B)** Plasmid map of pMOD-2<MCS>^{TM}. **(C)** Map of pMOD^{TM}-2<MCS> multiple cloning site flanked by ME and *Pvu*II sites. (Reproduced with kind permission from Epicentre Product Literature #145: EZ-Tn5^{TM} pMOD^{TM}-2<MCS> Transposon Construction Vectors, 2005. © 2005 Epicentre Technologies Corporation. All rights reserved.)

1. Transposon DNA is released by *Pvu* II digestion of plasmid DNA or PCR products.
2. Transposon DNA is gel-purified using the QIAquick gel extraction columns as recommended by the manufacturer.
3. Transposomes are preformed by incubating EZ::TN<KAN-2> transposon DNA (7 ng/µL; 8.9 nM) with hyperactive Tn*5* EZ::TN transposase (0.1 U/µL; 0.1 µM; Epicentre) in a solution containing 40 mM Tris-acetate (pH 7.5), 100 mM potassium glutamate, 0.1 mM EDTA, 1 mM dithiothreitol, and tRNA (0.1 mg/mL). Optimal transposase:DNA molar ratio is about 9:1 (**Note 6**).
4. Samples are incubated for 1 h at 37°C and dialyzed against the 10 mM Tris-acetate, pH 7.5, 1 mM EDTA buffer on 0.025-µm filters for 30 to 60 min.
5. Dialyzed samples are mixed with electrocompetent cells as described in **Section 3.2.4**.

3.2.4. Electroporation of E. coli *with Transposomes*

A low concentration of transposon DNA is required for efficient transposome formation. This results in large sample volumes for electroporation. Reaction mixtures can be concentrated prior to electroporation using the Microcon Centrifugal Filter Device YM-100 (Millipore). Alternatively, highly concentrated electrocompetent *E. coli* cells are prepared and mixed with dialyzed transposome reactions in a 2 : 1 (v/v) cells : transposome mix ratio. Cells are transformed using an Eppendorf electroporator 2510 at 2.4-kV field strength, 600-Ω resistance (fixed), and 10-µF capacitance (fixed). Cultures are immediately diluted with the medium prepared for outgrowth (**Section 3.3**) minus antibiotic and incubated at 37°C for 40 min with gentle agitation (150 rpm). Serial dilutions of recovered cells are plated to quantitate the total number of independent insertion mutants obtained. Recovered transformants can be stored as aliquots in 20% glycerol for future use (**Note 7**) or immediately subjected to outgrowth.

3.3. Outgrowth of the Mutagenized Population

Design of the outgrowth conditions is central to a genetic footprinting experiment and largely determines which genes will be identified as essential. Whereas inactivation of some genes will be lethal under any growth conditions ("essential for survival" genes), others will switch their essentiality depending on conditions during outgrowth (conditionally essential genes). Genes essential for survival have been traditionally assayed by testing for mutant colony formation on solid complex medium. However, many more growth conditions can be designed for analysis of gene essentiality, including propagation in various defined minimal or complex media (solid or liquid), growth under different stresses, survival in animal models of infection (for pathogenic *E. coli* strains), and so on. For example, we have determined a minimal set of genes required for *E. coli* aerobic logarithmic growth in complex medium with a special emphasis on essential genes in vitamin and cofactor biosynthetic pathways ***(14, 17)***. In this study, outgrowth of the mutagenized population was conducted in complex LB-based liquid media (**Note 8**) supplemented with vitamin and cofactor precursors that are readily salvaged by the cells:

1. After electroporation, the mutagenized population was inoculated in a BIOFLO 2000 fermentor into 900 mL of preheated media (described in **Section 2**).

2. Throughout the fermentation, temperature was held at 37°C, dissolved oxygen was held at 30% to 50% saturation, and the pH was held at 6.95 via titration with 5% H_3PO_4.
3. Cells were grown in batch culture for 23 population doublings (12 h) to a cell density of 1.4×10^9 (**Note 9** and **Chapter 24**). Genomic DNA was isolated and used to generate genetic footprints.

3.4. Footprint Generation: Detection of Surviving Mutants

3.4.1. Purification of Genomic DNA (gDNA)

Genomic DNA isolation kits are commercially available from Fermentas (Nahover, MD), Bio-Rad (Hercules, CA), BD Biosciences-Clontech (Palo Alto, CA), Qiagen (Valencia, CA), Sigma (St. Louis, MO), and other companies. An alternative large-scale genomic DNA isolation protocol is suggested below.

1. Using cell pellets from 500 mL of cell culture, add 40 mL of resuspension buffer (10 mM Tris HCl, pH 7.5, 1 mM EDTA, 1 µg/mL RNaseA) to resuspend the cells.
2. Add lysozyme to a final concentration of 2 to 5 mg/mL and incubate for 30 to 40 min in a water bath with occasional gentle mixing.
3. Add Proteinase K (150 to 200 µg/mL final concentration) plus SDS (0.5% to 2.0% final), and incubate for 1 to 2 h at 55°C. Incubate longer if necessary.
4. Add 0.5 volume of phenol; mix gently for at least 10 to 15 min.
5. Centrifuge at 10,000 rpm for 15 min at 15°C. It is important to use wide-bore pipette tips to reduce the shearing of genomic DNA.
6. Extract with phenol:chloroform mix (1 : 1) several times (3 to 5 times) until no white interphase is visible.
7. Extract once with 1 volume chloroform.
8. Add 1/10 volume of 3 M NaAc, mixing well before adding 1.5 volume of 100% ethanol. Mix gently and notice the strings of DNA precipitate out of solution.
9. Take out the “knot” of DNA with a pipette tip.
10. Gently rinse the DNA in 70% ethanol and place in a solution of 70% ethanol for storage.

3.4.2. Detection of Transposon Insertions Using Nested PCR

A nested PCR approach is useful in reducing artifacts in detection of transposon insertions. This approach utilizes two pairs of primers, which are used consecutively, with the second pair of primers nested within the first (**Fig. 2**). Each primer pair contains one universal transposon-specific primer and one chromosome-specific primer. Chromosome-specific primers can be designed as an ordered set of unidirectional primer pairs covering the entire *E. coli* genome or any given region of it (**Note 10**). Multiple strategies of positioning genomic primers can be envisioned. One approach uses primer pairs covering large, contiguous genomic regions separated by approximately 3500 bp (**Note 11** and **Fig. 2A**), while primers within each unidirectional pair are separated by the shortest possible distance in the range of −3 to 900 bp. An average primer is 27 nt long with annealing temperature of 68°C to 72°C. Transposon-specific primers are chosen to avoid any significant similarity with the *E. coli* chromosome, using PrimerSelect software (DNASTAR) or Web-based primer design software such as Primer3 or NetPrimer (**Section 2**).

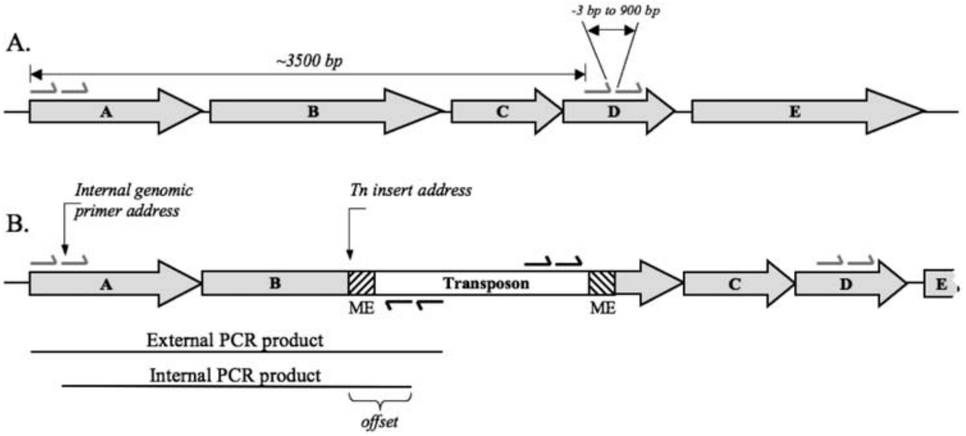


Fig. 2. Detection and mapping of transposition events by two-step nested PCR. **(A)** Primer design: genomic landmark primer pairs (shown as gray tandem arrows) are spaced on average by ~3500 bp, covering the entire genome, with –3 (overlap) to 900 bp between the primers in each pair. **(B)** Mapping transposition events by nested PCR. Each PCR reaction ("external" and "internal") utilizes one genomic (gray arrow) and one orientation-specific transposon (black arrow) primer with the second "internal" pair of primers nested within the first.

Two pairs of nested, outwardly directed transposon-specific primers (one at each end) are used to detect transposons inserted in both orientations (**Fig. 2B**). Using EZ::TN<KAN-2> as an example, the "forward" primer pair includes an external primer, 5′-GTTCCGTGGCAAAGCAAAAGTTCAA-3′, and an internal primer, 5′-GGTCCACCTACAACAAAGCTCTCATCA-3′. The "reverse" primer pair includes an external primer, 5′-CCGACATTATCGCGAGCCCATTTAT-3′, and an internal primer, 5′-GCAAGACGTTTCCCGTTGAATATGGC-3′. The composition of the PCR reaction mixtures is described in **Table 1** and cycling conditions are described in **Table 2**.

Table 1
Composition of the PCR Reaction Mixtures

Components	Amount in external PCR mix	Amount in internal PCR mix
DNA template	0.3 µg*	10^3-fold dilution of external PCR mix
dNTPs (10 mM each)	0.2 mM	0.2 mM
Advantage 10× reaction buffer	2.0 µL	2.0 µL
Genomic primer	0.4 mM external	0.4 mM internal
Transposon primer	0.4 mM external	0.4 mM internal
50× Advantage cDNA polymerase mix	0.4 µL	0.4 µL
Distilled H_2O	to 20 µL final volume	to 20 µL final volume

*Equivalent of 6×10^7 *E. coli* genomes.

Table 2
Touchdown PCR Reaction Conditions

Description	External PCR	Internal PCR
Hot start	95°C for 1 min	95°C for 1 min
Cycle 1	94°C for 12 s 70°C for 6 min	None
Number of cycles:	*2*	
Cycle 2	94°C for 12 s 69°C for 6 min	94°C for 12 s 69°C for 6 min
Number of cycles:	*2*	*2*
Cycle 3	94°C for 12 s 68°C for 6 min	94°C for 12 s 68°C for 6 min
Number of cycles:	*36*	*9*
Final extension	68°C for 6 min	68°C for 6 min

3.4.3. Agarose Gel Imaging and PCR Product Size Determination

PCR products from the second, nested, or "internal" reactions are size-separated and visualized on 0.65% agarose gels in the presence of a linear DNA ladder (**Note 12** and **Fig. 3A**). Gel electrophoresis should be optimized for DNA loading amounts and running time to obtain separation that is suitable for precise PCR product length determination. PCR products appear as discrete bands, and nondiscrete bands (fuzzy or low

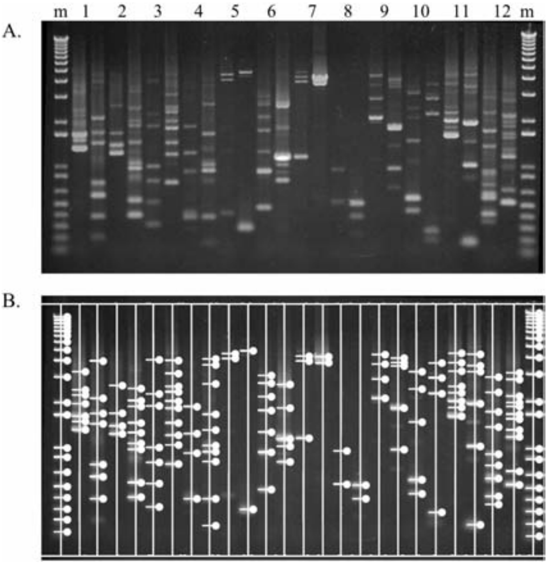


Fig. 3. Agarose gel electrophoresis and semiautomatic size determination of PCR products. **(A)** Internal PCR products are size-separated on 0.65% agarose gels in the presence of 1-kb linear DNA ladder ("m" lanes). **(B)** The Kodak 1D Gel Analysis Software is used to compare mobility of the PCR products in experimental lanes to that of the ladder for automatic size determination.

intensity when stained with ethidium bromide) may indeed be PCR artifacts. The sizes of discrete bands (in bp) are measured using software programs, available for gel image analysis, for example, from Labworks Software (UVP Inc., Upland, CA), Kodak 1D Image Analysis Software (Eastman Kodak, Rochester, NY), or other companies. Bands (PCR products) are called in the experimental gel lanes and compared with bands called in the molecular marker lane (**Fig. 3B**). These programs also generate text files that contain the information on a given lane or a gel.

3.4.4. Calculating Transposon Insertion Addresses

Each transposon insertion address is determined using a chromosomal location of the corresponding internal genomic primer and the offset of the 5′ position of the internal transposon-specific primer (**Fig. 3B**). The offset (fixed for a given Tn-specific primer pair) refers to the distance (in bp) between the Tn*5* mosaic end (ME) and the 5′ end of the internal primer (**Fig. 3B**). The insertion addresses can be calculated en masse in Excel as follows:

[Internal genomic primer address] + [Nested PCR product size] –
[Internal Tn primer offset] = Tn insert address

Table 3 illustrates these calculations for a sample genomic primer address (123,456 bp). Note that each genomic primer is used for detecting and mapping insertions of both transposon orientations within each ~3500-bp window.

3.4.5. Optimization of Experimental Conditions for Low-Noise Detection of Transposon Insertions

Reaction conditions can be optimized for detection of the maximum number of inserts while keeping the level of noise introduced by PCR low using a small set of

Table 3
Calculating Chromosomal Addresses of Transposon Insertions

Internal genomic primer address (bp)	Internal PCR product size (bp)	Internal Tn primer offset (bp)		Tn insert address (bp)
		Forward primer	Reverse primer	
123,456	200	100		123,556
123,456	351	100		123,707
123,456	690	100		124,046
123,456	1131	100		124,487
123,456	2345	100		125,701
123,456	205		167	123,494
123,456	233		167	123,522
123,456	709		167	123,998
123,456	1198		167	124,487
123,456	2479		167	125,768

control genes with known essentiality. First, the minimal amount of template genomic DNA that contains a representative mix of all mutant chromosomes and consistently yields reproducible patterns of bands in a PCR reaction should be determined (e.g., 0.3 μg of DNA is equivalent to 6 × 10^7 *E. coli* genomes). Second, the products of an external and the corresponding internal PCR reactions can be analyzed on an agarose gel side by side. This comparison is used as a guide for optimizing PCR parameters (such as cooling rate, number of cycles in external vs. internal PCR, and dilution rate of external PCR products) in order to minimize the number of false products, namely internal PCR bands lacking the corresponding external PCR product. Third, a number of internal PCR bands can be gel-purified and sequenced and the calculated insert locations compared with sequencing results. In our hands, the mapping error determined in this manner amounted to approximately ±4.5% of the size of each PCR band.

3.5. Genomic Mapping of Transposition Events and Determination of Gene Essentiality

3.5.1. Genomic Mapping of Transposition Events

Positions of transposon insertions calculated as described above can be mapped onto a chromosomal map either manually or semiautomatically using a software program to compare insert locations with addresses of protein-coding genes and nontranslated RNAs. An example output of a simple program developed in our group is presented in **Table 4**. While a simple table similar to the one shown in **Table 4** can be sufficient for making essentiality calls, a graphic output (*see* **Fig. 5A–D**) simplifies this task. However, more elaborate software is needed to support this feature.

3.5.2. Assessment of Conditional Gene Essentiality Based on Genetic Footprinting Data

Gene (non)essentiality assertions are made based on analysis of the number and relative positions of inserts within each gene after selective outgrowth as well as the relative intensity of electrophoresis bands corresponding with each transposition event. Retention of multiple inserts generally identifies a gene as dispensable under conditions tested. Failure to recover inserts or the presence of only a limited number of inserts at the very ends of a coding sequence suggests that a gene is essential under specific growth conditions (**Fig. 4**). Gene essentiality conclusions can be generated manually or semiautomatically. In the latter case, automatic calls should be manually confirmed or corrected at least for "borderline" cases (ORFs containing a single insert or with inserts close to 5′ or 3′ ends). A few guidelines on assertion of conditional gene essentiality are given below.

3.5.3. ORFs Asserted as Undetermined

ORFs are excluded from essentiality analysis if no reliable PCR data can be obtained for the corresponding region of the *E. coli* chromosome for technical reasons, such as PCR failure, nonspecific primer annealing in areas of DNA repeats, or insufficient length of generated PCR products. The latter case is illustrated in **Figure 5A**. This problem may be due to simultaneous synthesis of an excessively large number of PCR

Table 4
Example Output of the Transposition Mapping Software

	Well			Insert location relative		Band
ORF	number	Primer address	Calculated insert address	to (amino acids)	ORF	intensity
RE	**B0795**	**HAS BANDS**	**Periplasmic component efflux system**			
	E.03	Pr828620i	Escherichia_coli_K12_829,147	Outside ORF	—	0.076
	E.03	Pr838620i	Escherichia_coli_K12_829,111	Inside ORF	84	0.062
	E.03	Pr838620i	Escherichia_coli_K12_828,996	Inside ORF	199	0.196
	E.03	Pr838620i	Escherichia_coli_K12_828,709	Inside ORF	486	0.304
	E.02	Pr825051i	Escherichia_coli_K12_828,573	Inside ORF	622	0.069
REC0076	**RhlE**	**HAS BANDS**	**ATP-dependent RNA helicase**			
	E.03	Pr828620i	Escherichia_coli_K12_830,130	Inside ORF	35	0.272
	E.03	Pr828620i	Escherichia_coli_K12_830,167	Inside ORF	72	0.053
	E.03	Pr828620i	Escherichia_coli_K12_830,435	Inside ORF	340	0.134
REC0075	**B0791**	**NO BANDS**	**Hypothetical protein**			
REC04633	**YbiA**	**NO BANDS**	**Hypothetical cytosolic protein**			

Detected transposition events are organized and listed according to an ORF they have occurred at (or near). The first line in each group of entries is a summary containing an automatic assessment of an ORF's fate ("has bands," "no bands," or "not covered") as well as an ORF name, b-number, and predicted function. If any insertions have been mapped within an ORF, a detailed description of each transposition event is given, including a chromosome address and a well number of the landmark primer used in detection, a calculated insert address, insert location within the ORF (in amino acid residues from the start codon), and the relative intensity of the corresponding electrophoretic band.

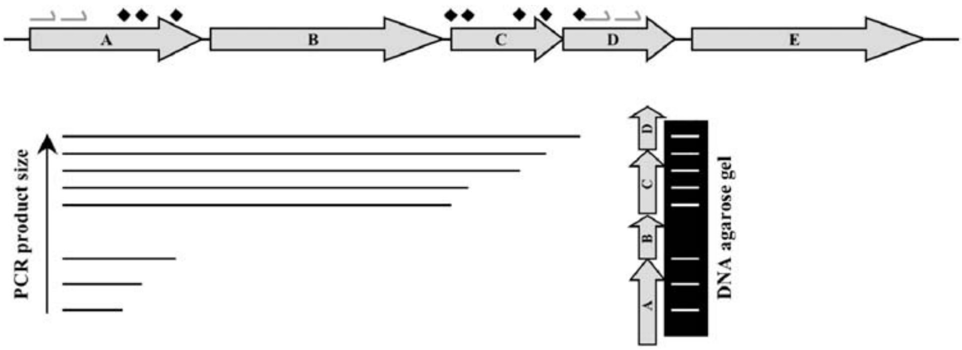


Fig. 4. Genetic footprinting for detection of essential and dispensable genes. Blank regions on electrophoresis gels correspond with DNA regions retaining no insertions after selective outgrowth, implicating the genes located in such loci as potentially essential. Diamonds indicate transposon insertion points as calculated by PCR product length.

products in a particular reaction, higher local GC content, stable secondary structure of PCR intermediates, and so on.

3.5.4. ORFs Asserted as Essential

All ORFs of sufficient length (**Note 13**) free of inserts can be asserted as essential if located in regions for which reliable PCR data were obtained. In addition, genes with only a few insertions within the 3′-most 20% or 5′-most 5% of the gene can be considered essential because in many cases these insertions were found to be nondisruptive *(2, 5, 14)*. Examples of essential genes (ORFs 4 through 7) are shown in **Figure 5A, B**. It should be kept in mind that the lack of insertions within an ORF may be due to reasons other than essentiality (for detailed discussion, see Refs. *14, 15, 34*), such as "cold spots" for transposition or polar effects, when insertions into a dispensable gene are selected against due to their disruptive effect on the transcription of a downstream essential gene.

3.5.5. ORFs Asserted as Nonessential (Dispensable) Under Experimental Conditions

A gene with one or more insertions located within 5% to 80% of its coding length should be considered nonessential (*see* ORFs 10, 14, and 15 in **Fig. 5C, D**), except for relatively long ORFs (>1000 bp). The latter are potentially essential genes with a few regions where insertion can be tolerated, especially if insertion density within the coding sequence is significantly below local average (examples are ORFs 11 and 12 in **Fig. 5D**). Correct essentiality assertions are notoriously difficult to make in such cases *(15, 34)*. Another likely source of erroneous "dispensable" assertions is proteins consisting of two or more independently functioning domains. For example, inserts may be tolerated within the 3′ region of a corresponding gene if the C-terminal domain of a protein is associated with a dispensable function. This may occur even when a function associated with the N-terminal domain is genuinely essential (as with FtsX *[14]*). Assertion of *ftsX* as nonessential will result in missing the essential function encoded by its 5′ portion.

3.5.6. ORFs Asserted as Ambiguous

ORFs were asserted as ambiguous if the experimental evidence was insufficient to make specific conclusions about essentiality. These cases are different from those ones classified as "undetermined," where no experimental data were obtained for technical reasons. Assertion failure in cases of "undetermined" ORFs may be technical in nature, and repeating PCR amplification or gel electrophoresis can improve the situation. In cases of ambiguous ORFs, assertion failures are likely due to underlying biological reasons and are harder to resolve. Examples of "ambiguous" essentiality assessments

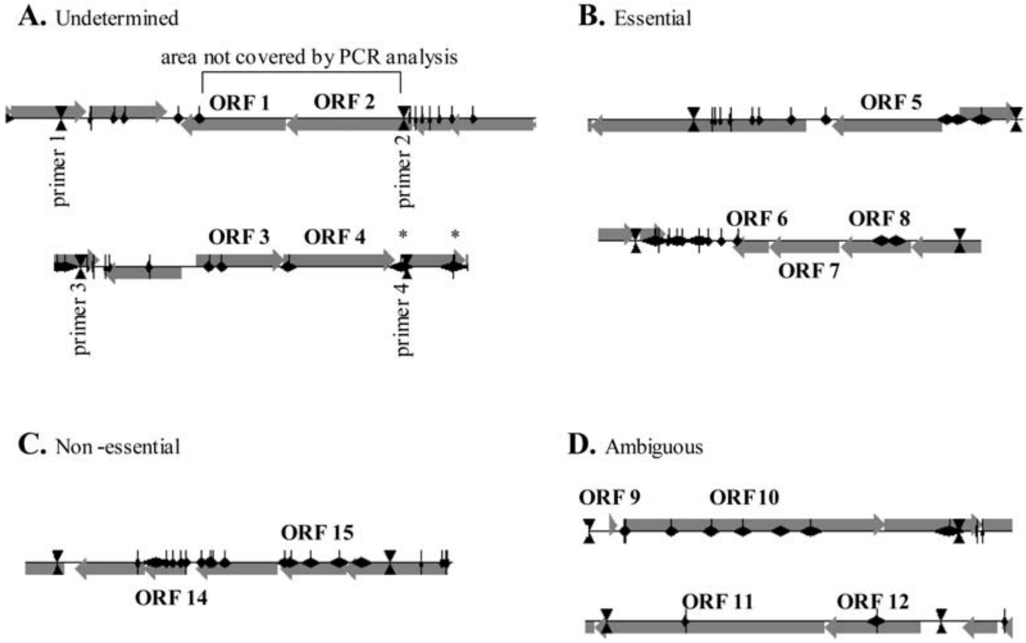


Fig. 5. Assessment of conditional gene essentiality based on genetic footprinting data. **(A–D)** Examples of graphic output of the transposition mapping software. In each panel, large horizontal arrows indicate the length and direction of predicted open reading frames. Positions of landmark PCR primers are shown by bows crossing the chromosome. Black diamonds represent detected transposon insertions. The width of each diamond corresponds with a mapping error introduced by gel electrophoresis (generally equal to ~4.5% of the size of each PCR product). Note that mapping errors (diamond widths) are small for inserts located in the immediate vicinity of each primer but grow as the distance from a primer increases. The vertical line associated with each diamond shows a relative intensity of the corresponding electrophoretic band. **(A)** ORFs 1 and 2 were asserted as undetermined due to insufficient length of PCR products obtained with landmark primer 1. The longest PCR product originating from primer 1 is significantly shorter than the distance between landmark primers 1 and 2. Conversely, the region between primers 3 and 4 was "covered" in its entirety because the longest PCR products (marked by asterisks) originating from primer 3 are longer than the distance between primers 3 and 4. **(B)** Examples of ORFs asserted as essential under experimental conditions. **(C)** Examples of ORFs asserted as dispensable under conditions tested. **(D)** ORFs 11 and 12 were asserted as ambiguous.

(**Fig. 5B, D**) may include (1) ORFs shorter than an average distance between transposition events (ORF 9), (2) relatively long ORFs with a single insertion (ORFs 11 and 12), and (3) ORFs containing only inserts corresponding with PCR products of low intensity (ORF 8).

The success rate of any high-throughput genome-scale project is never 100%. Unavoidably, a number of ORFs will lack informative essentiality assertions upon completion of the genetic footprinting procedure described here; however, this does not undermine the value of simultaneous determination (under uniform growth conditions) of conditional essentiality for the vast majority of *E. coli* ORFs that this approach can provide with a relatively low investment of time and money.

Notes

1. Epicentre Technologies (Madison, WI) carries the largest selection of Tn*5* and *Mu*-based transposition tools specialized for various purposes, including transposon construction vectors and transposon insertion kits. Bacteriophage *Mu*–based transposition products can also be purchased from (1) Finnzymes (Espoo, Finland): MuA Transposase, Entranceposons, transposon construction vectors, transposition kits; and (2) Invitrogen (Carlsbad, CA): GeneJumper Kits. Tn7-based reagents can be obtained from NEB (Beverly, MA): TnsABC Transposase and Transprimer kits.
2. For the *E. coli* K-12 MG1655 (with the genome size of 4,639,221 bp), the 10^5 independent mutants (generated and analyzed in ***[17]***) should result in the average insertion density of 1 per 46 bp. However, only 1.8×10^4 distinct insert locations were experimentally detected in the *E. coli* chromosome (average density of 1 insert per 258 bp). This discrepancy was apparently due to slight preferences in target sequence recognition by the modified Tn*5* transposase and can be used as a numerical measure of such bias. In this experiment, there was only ~sixfold difference between the observed insertion density and the density expected from completely random transposition. Increased insertion density around the *E. coli* chromosomal origin of replication (*oriC*) and slightly lower density near the terminus (*dif*) may have contributed to this bias ***(17)***. The mean insertion density of 1/258 bp corresponded with 3.9 inserts per ORF (for an average *E. coli* ORF size of ~1 kb), allowing unambiguous essentiality assessments for 87% of *E. coli* ORFs.
3. The Tn*5* transposase supplied by Epicentre is a hyperactive triple mutant of the wild-type enzyme. The introduced mutations (E54K, M56A, L372P) along with modifications of the 19-bp transposase recognition sequences have enhanced transposition efficiency approximately 1000-fold compared with that of the wild-type Tn*5* ***(23, 35)***.
4. Custom or artificial transposons designed with outward-directed promoters may alleviate potential polar effects on expression of downstream genes. However, we have found this to be unnecessary in *E. coli*. In numerous cases, viable EZ::TN<KAN-2> insertions were detected upstream of known essential genes, even though no specific promoter sequence was added to its structure ***(14)***. This apparently reflects the fact that the EZ::TN<KAN-2> element inserted in either orientation is capable of initiating a level of transcription sufficient for survival in many cases. Similar observations were made for the Ty1 transposon by Smith and co-workers ***(2)***.
5. Taq DNA polymerase adds 3′ A overhangs to a significant fraction of PCR products. The use of enzyme cocktails containing a proofreading thermostable DNA polymerase for PCR amplification of transposon DNA, although helpful, does not completely eliminate this problem.

6. Concentrations of the transposon DNA and transposase in the reaction mix as well as the transposase/DNA molar ratio have a tremendous effect on transposition rates and have to be carefully optimized for each new batch of DNA and transposase. Numbers given here should be used merely as a starting point for optimization. Reaction conditions should favor formation of monomer transposome complexes, when the two transposase subunits bound to the ME sequences of the *same* DNA molecule dimerize preferentially, causing DNA circularization via synaptic complex formation. Incubation at higher DNA concentrations leads to formation of multimers via dimerization of transposase subunits bound to different DNA molecules. Multimers are not active in electroporation and cause inherent loss of transposase.
7. A single cycle of freezing and thawing of *E. coli* cells results in a minimum five-fold reduction in titer, equivalent to 80% loss of mutagenized library. For this reason, we prefer to prepare a new population of insertion mutants for every experiment. However, it is recommended that a fraction of every mutagenized population is frozen and stored as a "time zero" control sample (prior to outgrowth).
8. Selective properties of liquid culture differ significantly from that of solid media. Growth of a complex mutant population in liquid may lead to rapid loss of nonlethal mutants with merely retarded growth rates due to competition with faster growing cells, whereas on solid media the same variants can be capable of forming colonies, albeit slower. This may give rise to "overcalling" essential genes. Conversely, cross-feeding in a complex mutant population, especially at higher cell densities, can complement potentially lethal mutations and may lead to missing essential ORFs. Nonetheless, we opted for selective outgrowth in liquid medium due to the fact that it allows genetic dissection of particular cellular processes and/or growth phases, unattainable during testing "for colony formation." For example, using strictly controlled growth in liquid media genes essential for logarithmic growth can be assayed independently of genes required for survival in stationary phase, or influence of such factors as partial oxygen pressure or pH on viability can be assayed. In addition, growth conditions within a colony (as opposed to liquid growth) are non-uniform, which may complicate interpretation of results. In any case, specifics of the outgrowth conditions should be considered during the interpretation of genetic footprint results.
9. The minimal number of population doublings necessary to reduce the titer of cells with insertions in core essential genes beyond detection level was determined in a pilot study. Longer outgrowth will lead to disappearance of mutants with less severe growth defects. Each mutant growth rate is inversely proportionate to duration of the outgrowth required for its complete disappearance from the population.
10. Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3.cgi) software was used to design a set of >1200 unidirectional primer pairs covering the entire *E. coli* genome. To simplify this task, a software program can be written to automate the input of *E. coli* genomic sequences into Primer3 and for sorting and formatting the output. The designed primers can also be arrayed in 96-well format, thus allowing robotics-assisted PCR and reagents preparation. Each primer pair is assigned a unique name (chromosomal address of the 5′ end of the internal landmark primer) and a plate/well number. Internal and external primers in each pair can be distinguished by an additional letter and grouped in different plates ("e" and "i" plates). To streamline a genome-scale footprinting project, most of the steps should be performed in the same 96-well format, starting with primer design and ordering and including PCR reactions, gel electrophoresis, and calculation of insert addresses.
11. The optimal distance between landmark primers is determined by an average density of insertions and by PCR reaction conditions. Simultaneous synthesis of an excessive number

of products in each PCR reaction sample may lead to reduced amplification of each specific product (especially longer ones), undermining their detection and analysis on electrophoretic gel images. In addition, polymerase cocktails used in various PCR kits differ by their ability to amplify DNA fragments longer than a few kilobases. In our experience, PCR products of no longer than 3500 bp in length could be reliably generated and detected (with 12 independent PCR products generated per reaction on average).

12. Useful linear DNA ladders should have multiple fragment sizes. Ideally, the low-molecular-weight fragments should have 50- to 100-bp gradations whereas larger fragments have ~1000-bp gradations. The 1-kb Plus DNA Ladder (Invitrogen) was used in published whole-genome footprinting studies *(14, 17)* for accurate size determination of PCR products.
13. Lack of insertions in relatively short ORFs (250 to 300 bp long), consistently interpreted here as an indication of gene essentiality, can in fact be accidental as the average density of detectable transposon insertion is 1 per 258 bp.

Acknowledgments

We are grateful to all our former colleagues at Integrated Genomics, Inc., for their tremendous help and advice during the whole-genome gene essentiality screening in *E. coli*. The authors would like to thank Andrei Osterman for his guidance and valuable comments on the manuscript and Michelle Rodriguez for excellent help with the figures.

References

1. Smith, V., Botstein, D., and Brown, P. O. (1995) Genetic footprinting: a genomic strategy for determining a gene’s function given its sequence. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 6479–6483.
2. Smith, V., Chou, K. N., Lashkari, D., Botstein, D., and Brown, P. O. (1996) Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* **274**, 2069–2074.
3. Hutchison, C. A., Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., et al. (1999) Global transposon mutagenesis and a minimal mycoplasma genome. *Science* **286**, 2165–2169.
4. Reich, K. A., Chovan, L., and Hessler, P. (1999) Genome scanning in *Haemophilus influenzae* for identification of essential genes. *J. Bacteriol.* **181**, 4961–4968.
5. Akerley, B. J., Rubin, E. J., Novick, V. L., Amaya, K., Judson, N., and Mekalanos, J. J. (2002) A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 966–971.
6. Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2001) Comprehensive identification of conditionally essential genes in mycobacteria. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 12712–12717.
7. Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* **48**, 77–84.
8. Sassetti, C. M., and Rubin, E. J. (2003) Genetic requirements for mycobacterial survival during infection. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12989–12994.
9. Wong, S. M., and Mekalanos, J. J. (2000) Genetic footprinting with *mariner*-based transposition in *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 10191–10196.

10. Liberati, N. T., Urbach, J. M., Miyata, S., Lee, D. G., Drenkard, E., Wu, G., et al. (2006) An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2833–2838.
11. Jenks, P. J., Chevalier, C., Ecobichon, C., and Labigne, A. (2001) Identification of non-essential *Helicobacter pylori* genes using random mutagenesis and loop amplification. *Res. Microbiol.* **152**, 725–734.
12. Kwon, Y. M., Kubena, L. F., Nisbet, D. J., and Ricke, S. C. (2003) Isolation of *Salmonella typhimurium* Tn*5* mutants defective for survival on egg shell surface using transposon footprinting. *J. Environ. Sci. Health B* **38**, 103–109.
13. Badarinarayana, V., Estep, P. W. 3rd, Shendure, J., Edwards, J., Tavazoie, S., Lam, F., and Church, G. M. (2001) Selection analyses of insertional mutants using subgenic-resolution arrays. *Nat. Biotechnol.* **19**, 1060–1065.
14. Gerdes, S. Y., Scholle, M. D., D'Souza, M., Bernal, A., Baev, M. V., Farrell, M., et al. 2002. From genetic footprinting to antimicrobial drug targets: examples in cofactor biosynthetic pathways. *J. Bacteriol.* **184**, 4555–4572.
15. Hare, R. S., Walker, S. S., Dorman, T. E., Greene, J. R., Guzman, L. M., Kenney, T. J., et al. (2001) Genetic footprinting in bacteria. *J. Bacteriol.* **183**, 1694–1706.
16. Winterberg, K. M., Luecke, J., Bruegl, A. S., and Reznikoff, W. S. (2005) Phenotypic screening of *Escherichia coli* K-12 Tn*5* insertion libraries, using whole-genome oligonucleotide microarrays. *Appl. Environ. Microbiol*. **71**, 451–459.
17. Gerdes, S., Scholle, M., Campbell, J., Balazsi, G., Ravasz, E., Daugherty, M., et al. (2003). Experimental determination and system-level analysis of essential genes in *E. coli* MG1655. *J. Bacteriol.* **185**, 5673–5684.
18. Hamer, L., DeZwaan, T. M., Montenegro-Chamorro, M. V., Frank, S. A., and Hamer, J. E. (2001) Recent advances in large-scale transposon mutagenesis. *Curr. Opin. Chem. Biol.* **5**, 67–73.
19. Hayes, F. (2003) Transposon-based strategies for microbial functional genomics and proteomics. *Annu. Rev. Genet.* **37**, 3–29.
20. Haapa, S., Taira, S., Heikkinen, E., and Savilahti, E. (1999) An efficient and accurate integration of mini-Mu transposons in vitro: a general methodology for functional genetic analysis and molecular biology applications. *Nucleic Acids Res.* **27**, 2777–2784.
21. Maekawa, T., Yanagihara, K., and Ohtsubo, E. (1996) A cell-free system of Tn*3* transposition and transposition immunity. *Genes Cells* **1**, 1007–1016.
22. Maekawa, T., Yanagihara, K., and Ohtsubo, E. (1996) Specific nicking at the 3′ ends of the terminal inverted repeat sequences in transposon Tn*3* by transposase and an *E. coli* protein ACP. *Genes Cells* **1**, 1017–1030.
23. Goryshin, I. Y., and Reznikoff, W. S. (1998) Tn*5 in vitro* transposition. *J. Biol. Chem.* **273**, 7367–7374.
24. Reznikoff, W. S., Goryshin, I. Y., and Jendrisak, J. J. (2004) Tn*5* as a molecular genetics tool: In vitro transposition and the coupling of in vitro technologies with in vivo transposition. *Methods Mol. Biol.* **260**, 83–96.
25. Bainton, R. J., Kubo, K. M., Feng, J. N., and Craig, N. L. (1993) Tn*7* transposition: target DNA recognition is mediated by multiple Tn*7*-encoded proteins in a purified *in vitro* system. *Cell* **72**, 931–943.
26. Chalmers, R. M., and Kleckner, N. (1994) Tn*10*/IS10 transposase purification, activation, and *in vitro* reaction. *J. Biol. Chem.* **269**, 8029–8035.
27. Griffin, T. J. T., Parsons, L., Leschziner, A. E., DeVost, J., Derbyshire, K. M., and Grindley, N. D. (1999) *In vitro* transposition of Tn*552*: a tool for DNA sequencing and mutagenesis. *Nucleic Acids Res.* **27**, 3859–3865.

28. Devine, S. E., and Boeke, J. D. (1994) Efficient integration of artificial transposons into plasmid targets *in vitro*: a useful tool for DNA mapping, sequencing and genetic analysis. *Nucleic Acids Res.* **22**, 3765–3772.
29. Lampe, D. J., Akerley, B. J., Rubin, E. J., Mekalanos, J. J., and Robertson, H. M. (1999) Hyperactive transposase mutants of the Himar1 *mariner* transposon. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 11428–11433.
30. Hoffman, L. M., Jendrisak, J. J., Meis, R. J., Goryshin, I. Y., and Reznikoff, W. S. (2000) Transposome insertional mutagenesis and direct sequencing of microbial genomes. *Genetica* **108**, 19–24.
31. Goryshin, I. Y., Jendrisak, J., Hoffman, L. M., Meis, R., and Reznikoff, W. S. (2000) Insertional transposon mutagenesis by electroporation of released Tn*5* transposition complexes. *Nat. Biotechnol.* **18**, 97–100.
32. Butterfield, Y. S., Marra, M. A., Asano, J. K., Chan, S. Y., Guin, R., Krzywinski, M. I., et al. (2002) An efficient strategy for large-scale high-throughput transposon-mediated sequencing of cDNA clones. *Nucleic Acids Res.* **30**, 2460–2468.
33. Bhasin, A., Goryshin, I. Y., Steiniger-White, M., York, D., and Reznikoff, W. S. (2000) Characterization of a Tn*5* pre-cleavage synaptic complex. *J. Mol. Biol.* **302**, 49–63.
34. Akerley, B. J., Rubin, E. J., Camilli, A., Lampe, D. J., Robertson, H. M., and Mekalanos, J. J. (1998) Systematic identification of essential genes by *in vitro mariner* mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 8927–8932.
35. Zhou, M., Bhasin, A., and Reznikoff, W. S. (1998) Molecular genetic analysis of transposase-end DNA sequence recognition: cooperativity of three adjacent base-pairs in specific interaction with a mutant Tn*5* transposase. *J. Mol. Biol.* **276**, 913–925.

7

Generating a Collection of Insertion Mutations in the *Staphylococcus aureus* Genome Using *bursa aurealis*

Taeok Bae, Elizabeth M. Glass, Olaf Schneewind, and Dominique Missiakas

Summary

Staphylococcus aureus is the leading cause of wound and hospital-acquired infections. The emergence of strains with resistance to all antibiotics has created a serious public health problem. Transposon-based mutagenesis can be used to generate libraries of mutants and to query genomes for factors involved in nonessential pathways, such as virulence and antibiotic resistance. Ideally, such studies should employ defined and complete sets of isogenic mutants and should be conducted so as to permit acquisition and comparison of the complete data sets. Such systematic knowledge can reveal entire pathways and can be exploited for the rational design of therapies. The *mariner*-based transposon, *bursa aurealis*, can be used to generate random libraries of mutants in laboratory strains and clinical isolates of *S. aureus*. This chapter describes a procedure for isolating mutants and mapping the insertion sites on the chromosome.

Key Words: *bursa aurealis*; *Himar 1* transposase; *mariner*; mutagenesis; *Staphylococcus aureus*; transposon library; temperature sensitive.

1. Introduction

The 2.7- to 2.9-Mbp genomes of several different *Staphylococcus aureus* strains have been sequenced, revealing large variability in size and gene content. The staphylococcal genomes encode between 2550 and 2870 genes (*1–3*). Over the past several decades, reverse genetic approaches have often been utilized to identify and characterize metabolic and biosynthetic pathways as well as virulence factors such as secreted toxins, surface proteins, or regulatory factors (*4–6*). Allelic replacement has been used extensively in this reverse genetic approach to generate mutations in chromosomal genes (*7*). To achieve this goal, mutated alleles of target genes are cloned into plasmids carrying replication-defective conditional mutations. Often, the mutated alleles correspond with gene deletions, frameshifts, or insertions of antibiotic resistance cassette. Under nonpermissive conditions, such plasmids integrate into the chromosome via

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

homologous recombination yielding merodiploid cells harboring both wild-type and mutant alleles (*8*). Plasmid resolution is achieved by growing cells under permissive conditions (*8*). Without markers for counterselection of the plasmid, allelic replacement with plasmid loss can be a very rare event that involves extensive screening and often several weeks or months of work (*9*). Thus, inserting an antibiotic marker in target genes is recommended for screening purposes (*8*). Transposon mutagenesis has been very useful in isolating larger collections of mutants. Tn*917* and signature-tagged mutagenesis were used to identify staphylococcal virulence factors (*10, 11*). Libraries of 1248 or 1520 randomly chosen (nonsequenced) transposon insertion mutants of *S. aureus* were analyzed in animal infections with mixed populations to reveal a competitive disadvantage of individual variants. Recently, we isolated 960 mutant variants with transposon Tn*917* and 10,325 with *bursa aurealis*. The sites of individual transposition events were examined by inverse polymerase chain reaction (PCR). This analysis revealed that, as expected, insertion of *bursa aurealis* into target DNA generates TA duplications at the insertion site, but unlike Tn*917*, *bursa aurealis* does not exhibit sequence preference in the genome (*12*). This analysis also suggested that Tn*917*-based libraries of 1248 or 1520 mutants examine only about 20% of the genes in the staphylococcal genome. Gene functions of *S. aureus* have also been examined using antisense RNA technology (Refs. *13* and *14* and **Chapters 19** and **20**). By cloning gene fragments in reversed orientation under control of an inducible promoter, the ability of antisense RNA sequences to interfere with *S. aureus* growth on agar plates was used to identify genes essential under these conditions. Two independent studies identified a total of 350 essential genes with a 30% overlap (*13, 14*). However, 110 of these presumed essential genes could be disrupted by *bursa aurealis*, suggesting that many of these assignments may not be correct (*12*). Although *bursa aurealis* does not allow the identification of essential genes, its ability to insert randomly allows for extensive and exhaustive studies of staphylococci biology.

2. Materials

1. pBursa, the transposon encoding plasmid.
2. pFA545, the transposase encoding plasmid.
3. *S. aureus* strains RN4220 and Newman.
4. Tryptic soy broth (TSB) and tryptic soy agar (TSA).
5. Chloramphenicol, erythromycin, and tetracycline antibiotics used at the following final concentrations 5, 10, and 2.5 µg/mL, respectively.
6. Incubators for plates and liquid cultures (30°C, 37°C, 43°C).
7. Centrifuge and microcentrifuge.
8. Dry ice/ethanol bath.
9. Electroporation equipment.
10. Lysostaphin (AMBI Products LLC, Lawrence, NY): The stock solution is prepared as a 2 mg/mL in 20 mM sodium acetate, pH 4.5, and kept frozen at −80°C or at 4°C for 4 weeks. The working solution is prepared in TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) and contains lysostaphin at a final concentration of 0.1 mg/mL.
11. Sterile freezing solution for long-term storage of bacterial strains at 80°C: 5% monosodium glutamate, 5% bovine serum albumin.
12. TSM buffer: 50 mM Tris-HCl pH 7.5, 0.5 M sucrose, 10 mM $MgCl_2$.

13. RNase (Sigma, Saint Louis, MO): Kept as a solution at 4 mg/mL in water, kept at 4°C for 4 weeks.
14. Agarose gel electrophoresis equipment.
15. Wizard Genomic DNA purification Kit (Promega, Madison, WI).
16. Oligonucleotide primers Martn-F (5′-TTT ATG GTA CCA TT CAT TTT CCT GCT TTT TC) and Martn-ermR (5′-AAA CTG ATT TTT AGT AAA CAG TTG ACG ATA TTC).
17. Restriction enzyme *Aci* I and T4 DNA ligase (New England Biolabs, Ipswich, MA).
18. QIAprep Spin Miniprep Kit, and PCR purification kit or MinElute 96 UF PCR purification kit (Qiagen, Valencia, CA).
19. PCR equipment.
20. Taq polymerase and buffer provided by the manufacturer (Promega, Madison, WI).

3. Methods

The methods described below outline (1) the generation of a *S. aureus* clinical isolate transformed with plasmids pBursa and pFA545, (2) the transposon mutagenesis, (3) the determination of insertion sites by inverse PCR, and (4) the identification of matching DNA sequences in the GeneBank.

3.1. Transformation of S. aureus Strain Newman with Plasmids pBursa and pFA545

The transposable element or transposon is encoded on plasmid pBursa (**Fig. 1**). The transposase is encoded on a second plasmid pFA545 (**Fig. 2**). The complete sequences

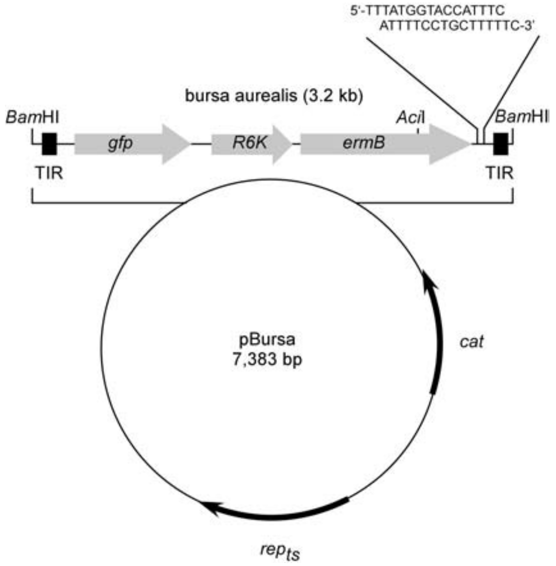


Fig. 1. Map of plasmid pBursa. *Bursa aurealis*, a mini-*mariner* transposable element, was cloned into pTS2, with a temperature-sensitive plasmid replicon (*rep*$_{ts}$) and chloramphenicol-resistance gene *cat* to generate pBursa. *Bursa aurealis* encompasses *mariner* terminal inverted repeats (TIR), green fluorescent protein gene (*gfp*), R6K replication origin (*oriV*), and erythromycin-resistance determinant *ermC*, an rRNA methylase. The positions of restriction enzymes recognition sites (*Aci* I and BamH I) are indicated.

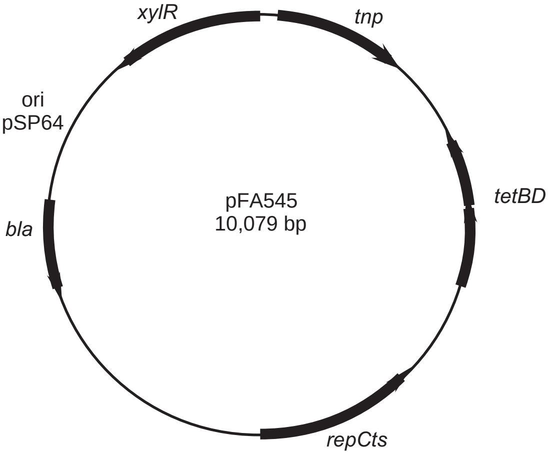


Fig. 2. Map of plasmid pFA545. Plasmid pFA545 is a derivative of pSPT181, a shuttle vector consisting of pSP64 with ampicillin resistance (*bla*) for replication and selection in *E. coli*, and pRN8103, a temperature-sensitive derivative of pT181 (*repCts*) and tetracycline-resistance marker (*tetB tetD*). The presence of *repCts* and *tetBD* allows for replication of pFA545 in *S. aureus* and other Gram-positive bacteria. *tnp*, *mariner* transposases; *xylR*, xylose repressor.

for both plasmids are available from the GeneBank (accession numbers AY672109 and AY672108).

3.1.1. Plasmid pBursa

The transposable element referred to as *bursa aurealis* is derived from the *Himar 1* (*mariner*) transposon. *Bursa aurealis* encompasses short inverted repeats of the horn fly transposon *(15, 16)*, the *ermC* resistance marker *(17)*, the R6K replication origin (*oriV*), and a promoterless *Aequorea victoria* green fluorescent protein (*gfp*) gene (**Fig. 1**). Insertion of *bursa aurealis* into *S. aureus* chromosome confers resistance to erythromycin and results in *gfp* expression if insertion occurs immediately downstream of a promoter. In principle, the presence of R6K replication origin allows rescue of transposon inserts along with the adjacent DNA fragments via cloning in *Escherichia coli* using a λ*pir* Tn*10* background that allows replication of R6K-based replicons. Unfortunately, selecting for erythromycin resistance in *E. coli* is not always possible due to high intrinsic resistance of most laboratory strains. Hence, this approach has not been exploited by our laboratory. *Bursa aurealis* was cloned into the pTS2 vector *(18)*, thereby generating pBursa (**Fig. 1**). pTS2 carries a temperature-sensitive replicon (pE194ts) and chloramphenicol resistance gene *(19, 20)*, allowing pBursa to replicate in most Gram-positive hosts. Staphylococcal cells bearing plasmid pBursa can be selected on chloramphenicol and erythromycin containing media at 30°C.

3.1.2. Plasmid pFA545

Plasmid pFA545 is a derivative of vector pSPT181 *(22)* and encodes the *Himar 1* transposase *(16)* cloned under the control of xylose-inducible *xylA* promoter and XylR repressor from *Staphylococcus xylosus* *(21)* (**Fig. 2**). The *xylA* promoter region was obtained from plasmid pIK64 ***(23)***. The parent vector pSPT181 is a shuttle vector consisting of pSP64, a ColE1-based replicon that can replicate in *E. coli* and contains ampicillin-resistance marker ***(24)***, and pRN8103 ***(25)***, a temperature-sensitive derivative of pT181 ***(26)*** that replicates in Gram-positive bacteria and carries the tetracycline resistance marker.

3.1.3. Electroporation of the Plasmids into S. aureus *RN4220*

The mutagenesis procedure is described for the clinical isolate Newman ***(27)*** but can be adapted to other strains as well. Due to the host restriction-modification system, pBursa and pFA545 plasmid DNA cannot be introduced into *S. aureus* Newman directly but need to be passaged through the laboratory strain RN4220. RN4220 is a vital intermediate for laboratory *S. aureus* manipulations, as it can accept *E. coli*–propagated plasmid DNA due to nitrosoguanidine-induced mutation(s) in its restriction-modification system ***(28)***, recently mapped to the *sau1hsdR* gene ***(29)***. Plasmids extracted from strain RN4220 can be electroporated in most staphylococcal isolates and other Gram-positive bacteria. The protocol below describes a method to electroporate plasmids pBursa and pFA545 extracted from strain RN4220 in strain Newman.

1. Streak *S. aureus* strain Newman from a frozen stock on a TSA plate and incubate overnight at 37°C.
2. Pick an isolated colony with a sterile loop and inoculate 2 mL TSB in a 100-mL flask.
3. Incubate overnight at 37°C with shaking.
4. Transfer the overnight culture in a 2-L flask containing 200 mL TSB.
5. Grow cells to mid-log phage ($OD_{600} = 0.5$) with vigorous shaking (approximate incubation time 2.5 to 3 h).
6. Transfer the culture into sterile spin bottles and collect cells by centrifugation at 5000 × *g* for 15 min.
7. Discard the supernatant and suspend the cell pellet in 40 mL of ice-cold sterile 0.5 M sucrose in deionized water.
8. Transfer the cell suspension to a prechilled 50-mL sterile centrifuge tube and keep on ice.
9. Collect cells by centrifugation at 8000 × *g*, 10 min, 4°C.
10. Discard supernatant and suspend the cell pellet in 20 mL of the ice-cold 0.5 M sucrose solution as above.
11. Collect cells by centrifugation at 8000 × *g*, 10 min, 4°C.
12. Repeat **steps 10** and **11** once more.
13. Resuspend the pellet in 2 mL ice-cold 0.5 M sucrose solution.
14. Transfer 100-μL aliquots of the prepared electrocompetent cells into microcentrifuge tubes chilled on ice.
15. Freeze tubes by plunging them in a dry ice-ethanol bath and store cells at –80°C until use (this protocol can be adapted to prepare larger volumes of competent cells).
16. For electroporation of pFA545, retrieve a tube of competent cells from the freezer and place tube on ice.
17. When cells are thawed, add 100 to 500 ng of purified plasmid.
18. Transfer the cell and DNA mix into a 0.1-cm prechilled electroporation cuvette (Bio-Rad, Hercules, CA).

19. Use the following settings for electroporation: voltage = 2.5 kV, resistance = 100 Ω, capacity = 25 µF.
20. Immediately after the pulse, add 1 mL TSB kept at room temperature and transfer entire contents to sterile Eppendorf tube.
21. Incubate for an hour at 30°C (no shaking required).
22. Pellet cells in a microcentrifuge (8000 × *g,* 3 min, RT) and remove most of the supernatant by flipping the tube upside-down.
23. Suspend cell pellet in remaining medium (50 to 100 µL) and spread cells on a TSA plate containing 2.5 µg/mL tetracycline ($\mathrm{TSA}_{\mathrm{tet2.5}}$).
24. Incubate plate at 30°C for at least 16 h (or until colonies are visible).

3.1.4. Isolation of Plasmid DNA from S. aureus Newman

25. Pick isolated colonies and grow cells in 5 mL $\mathrm{TSB}_{\mathrm{tet\ 2.5}}$ overnight at 30°C.
26. Transfer 1.5 mL of the overnight culture and collect cells by centrifugation (5000 × *g,* 3 min, RT); keep the remaining 3.5 mL cell culture at RT.
27. Suspend cell pellet in 50 µL TSM buffer.
28. Add 2.5 µL lysostaphin solution (2 mg/mL stock) and incubate for 15 min at 37°C (this will yield protoplasts).
29. Collect protoplasts by centrifugation (8000 × *g,* 5 min, RT) and discard supernatant.
30. Extract plasmid DNA from protoplasts using a QIAprep Spin Miniprep Kit (Qiagen) following the manufacturer's recommendations.
31. Analyze extracted plasmid by agarose gel electrophoresis (this procedure ensures that the plasmid has been successfully transformed into RN4220).
32. If the plasmid DNA is indeed present, the remaining cell culture can be kept frozen at −80°C in 50% sterile freezing solution.

3.1.5. Electroporation of the Passaged Plasmid DNA into S. aureus Newman

33. For electroporation of pBursa into *S. aureus* Newman, generate competent cells from the cells carrying pFA545 by repeating **steps 1** to **19** (**Section 3.1.3**).
34. Immediately after electroporation, spread cells on a TSA plate containing 2.5 µg/mL tetracycline and 2.5 µg/mL chloramphenicol ($\mathrm{TSA}_{\mathrm{tet2.5\ chl\ 5}}$); preincubation at 30°C is not necessary for the chloramphenicol or erythromycin selection.
35. Incubate plate at 30°C for at least 16 h (or until colonies are visible).
36. Repeat **steps 25** to **32** (**Section 3.1.4**) to verify the transformation.

Once *S. aureus* Newman or a strain of choice has been transformed with both plasmids, it is recommended to grow and freeze multiple isolates at −80°C as described in **step 32** (**Note 1**).

3.2. Transposon Mutagenesis

One of the main problems in generating a transposon mutant library is a potential disproportionate amplification of cells carrying the same transposon insertion. To minimize this unwanted process, mutants are isolated on solid medium. Nevertheless, the use of liquid culture remains an acceptable and a more rapid alternative.

1. Day 1: Streak the strain Newman carrying both plasmids, pBursa and pFA545, from frozen stock on $\text{TSA}_{\text{tet2.5 chl 5}}$ and incubate overnight at 30°C.
2. Day 2: Pick an isolated colony and inoculate 5 mL $\text{TSB}_{\text{tet2.5 chl 5}}$ (alternatively, use freshly transformed Newman for this step).
3. Incubate the cultures overnight at 30°C with shaking.
4. Day 3: Dilute the overnight culture 10^5-fold with sterile water and spread 50 to 100 µL of the diluted culture on one or more $\text{TSA}_{\text{tet2.5 chl 5}}$ plates.
5. Incubate overnight at 30°C.
6. Place a flask with 100 mL sterile water in 43°C incubator to preheat for the next day.
7. Day 4: Prewarm 10 to 20 $\text{TSA}_{\text{erm10}}$ plates at 43°C for 1 h.
8. Add 100 µL of sterile prewarmed water to 10 to 20 sterile microcentrifuge tubes (the number should be the same as for the number of plates in **step 7**).
9. Pick a colony from $\text{TSB}_{\text{tet2.5 chl 5}}$ plate (**step 5**) and mix with water in one of the microcentrifuge tubes. Vortex and repeat this procedure as needed.
10. Transfer 1 to 2 µL of cell suspension (**step 9**) and 100 µL of prewarmed water (**step 6**) onto a prewarmed $\text{TSA}_{\text{erm10}}$ plate (**step 7**).
11. Add 7 to 15 sterile glass beads on the plate, shake to spread cells evenly, remove glass beads, and collect in a separate container by inverting the plate.
12. Place plate immediately in a 43°C incubator and incubate until colonies appear (up to 2 days).
13. Inoculate colonies in 5 mL $\text{TSB}_{\text{erm10}}$ and incubate at 43°C overnight with shaking.
14. Freeze aliquots at −80°C in 50% sterile freezing solution. Use the remaining cells in each culture to map transposon insertion site(s) (see below).

Bursa aurealis transposition occurs at the frequency of ~10^{-6}. The system was meant to be inducible by design (**Fig. 2**); however, addition of xylose does not improve efficiency of transposition and is usually omitted. A typical experiment described above yields about 50 colonies per plate (**Section 3.2, step 12**). When more colonies (>200) appear, plasmid integration (**Note 2**) or incomplete loss of plasmid are generally suspected.

Liquid culture inoculation and incubation at 43°C (**step 13**) is performed for individual colonies isolated from **step 12**. Each of these isolates can then be subjected to inverse PCR and DNA sequencing. After purification of genomic DNA, it is observed that approximately 0.5% of all isolates fail to lose the plasmids (**Notes 3** and **4**).

If the investigator does not wish to sequence the sites of transposon insertions, isolated colonies (**step 12**) may be grown as pools (**step 13**).

A similar protocol can be used to isolate mutants in other Gram-positive bacteria (**Note 5** *[32]*).

3.3. Determination of Transposon Insertion Sites by Inverse PCR and DNA Sequencing

This step is not required if an investigator wishes to isolate a random (nonordered) library of mutants. However, it is highly recommended to sample a number of isolated mutant strains for quality control as the work proceeds. Sampling is described below for the analysis of 96 strains but can be adapted to a smaller sample size.

3.3.1. Purification of Chromosomal DNA with Wizard Genomic DNA Purification Kit

The following method is modified from the protocol of the manufacturer (Promega). Nuclei Lysis Solution and Protein Precipitation Solution are purchased from Promega.

1. Collect cells from 1.5 to 3 mL culture (**step 14**, **Section 3.2**) in an Eppendorf tube by centrifugation (8000 × *g,* 5 min, RT).
2. Discard the supernatant and suspend cell pellet in 50 µL TE buffer containing lysostaphin.
3. Incubate for 15 min at 37°C. The cell suspension will become viscous.
4. Add 300 µL of Nuclei Lysis Solution, vortex tubes, and transfer them to a heating block set at 80°C for 10 min. Longer incubations do not affect the quality of the DNA.
5. After incubation, the sample should become clear. In cases where this does not occur, pipette insoluble material up and down until the sample is clear. This treatment shears the DNA but does not affect the performance of inverse PCR.
6. Cool samples to room temperature and add 1.5 µL of RNase Solution (4 mg/mL in water), vortex briefly, and incubate 30 min at 37°C. Longer incubations do not affect the quality of the DNA.
7. Add 100 µL of Protein Precipitation Solution, vortex, and incubate on ice for 5 min.
8. Transfer tubes to a microcentrifuge and spin at top speed (16,000 × *g*) for 5 min.
9. Transfer supernatant to a clean microcentrifuge tube containing 300 µL of room-temperature isopropanol (you may need to repeat **step 8** once more if the sample is cloudy).
10. Vortex briefly, transfer tubes to a microcentrifuge, and spin at top speed (16,000 × *g*) for 5 min.
11. A DNA pellet should be visible at the bottom of the tube; discard supernatant by inverting tubes, add 750 µL of 70% ethanol kept at room temperature, and vortex briefly.
12. Transfer tubes to a microcentrifuge and spin at top speed (16,000 × *g*) for 5 min.
13. Remove remaining supernatant and dry pellets completely at room temperature.
14. Add 15 to 20 µL TE and rehydrate the DNA by incubating for 1 h at 65°C or overnight at 4°C.

3.3.2. Inverse PCR

In order to achieve a successful inverse PCR, two critical factors should be considered: (1) the choice of restriction site and enzyme used to digest genomic DNA and (2) the design of primers for amplification. Restriction enzymes recognizing DNA palindromes of six nucleotides or more generally generate DNA fragments that are too large to be amplified. On the other hand, the use of four nucleotide-based palindromes may yield fragments that are too small and do not permit unambiguous identification of transposon insertion sites. The four-nucleotide restriction site (CCGC) recognized by the *Aci* I enzyme is used here. This choice was driven by the knowledge that the *S. aureus* genome displays a rather low GC content (32%). Therefore, the average size of digestion products should in most cases be larger than the predicted size (4^4 = 256 bp). This is indeed the case (**Fig. 3**). To determine the appropriate set of primers, four primers were tested in four possible combinations. The optimal annealing temperature is best determined by using a **thermocycler** unit with **temperature gradient** capability. Martn-F and Martn-erm-R are the two oligonucleotides used for inverse

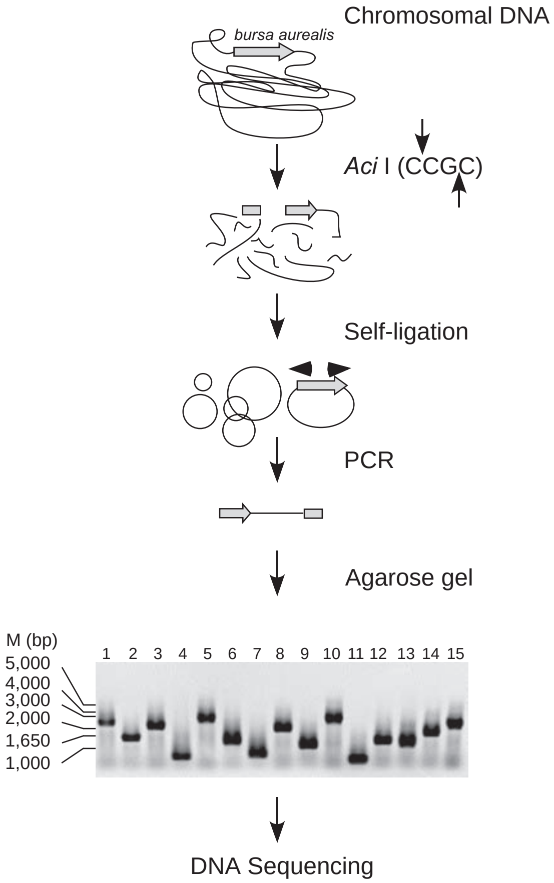


Fig. 3. Mapping insertion sites by inverse PCR. Genome DNA from 15 mutant strains of *S. aureus* Newman obtained by *bursa aurealis* transposon mutagenesis is isolated and digested with *Aci* I. Next, fragment self-ligation, inverse PCR, and agarose gel electrophoresis are performed. M indicates the molecular-weight marker (1-kb DNA ladder).

PCR (*see* sequences in **Section 2**). A diagram of the inverse PCR procedure is depicted in **Figure 3**.

1. Bring purified chromosomal DNA to room temperature or warm to 37°C to 65°C (this will help decrease viscosity of the sample).
2. Prepare reaction mix for digestion of DNA with *Aci* I in a 96-well plate assay:
   100 µL 10× buffer for *Aci* I (New England Biolabs)
   75 µL *Aci* I (New England Biolabs)
   325 µL water

Transfer 5 µL of the mixture to each well, add 5 µL chromosomal DNA.

3. Incubate samples 1 h and up to overnight at 37°C.
4. Inactivate *Aci* I by incubating samples for 20 min at 65°C.
5. Prepare ligation mix for the 96-well plate assay:

7.9 mL water
1.0 mL 10× buffer for T4 DNA ligase (New England Biolabs)
0.1 mL T4 DNA ligase (New England Biolabs)

Add to each well 90 µL of this mixture.

6. Incubate ligation reactions 3 h or up to overnight at room temperature.
7. Purify DNA with the Qiagen MinElute 96 UF PCR purification Kit (for 96-well sample) according to the manufacturer's protocol. Elute DNA in each well with 60 to 75 µL of elution buffer or deionized water.
8. Use 5 µL of ligated DNA for PCR reaction in a 25-µL reaction volume (Taq enzyme and 10× buffer from Promega are recommended; use primers at 1 µM each).
9. For primers Martn-F and Martn-ermR, the following 40-cycle program is recommended:
   30 s at 94°C
   30 s at 63°C
   3 min at 72°C.
10. Analyze 3 µL of the PCR reaction by 1% agarose gel electrophoresis (**Fig. 3**).
11. The DNA sequence of the PCR product is determined using the PCR product as a template and primer Martn-F.

3.4. Identification of Matching DNA Sequences in the GeneBank

Once the DNA sequence is determined, the identification of matching sequences in the GeneBank is trivial. However, the analysis of hundreds of sequences is cumbersome, and an automated method, like the one described here, can be used.

Step 1. DNA sequence files provided by the sequencing facility are configured into FASTA format and filtered for the transposon sequence prior to use in BLAST (*30*).

(a) FASTA format starts with a single-line description, followed by lines of sequence data. The description line is differentiated from the sequence data by a greater-than (">") symbol in the first column. The word following the ">" symbol is the identifier of the sequence, and the rest of the line is the description (both are optional).

(b) For each sequence file:
- Special characters are removed using the Unix command "tr -d '\r".
- The file name is used as the identifier and description of the DNA sequence.
- The transposon sequence substring CCTGTTA, indicating the end of the transposon, is searched for in the DNA sequence file.
- If found, the substring with an additional 160 nucleotides is extracted and combined with identifier and description of files.
- If the transposon sequence is not found, the first 260 nucleotides are extracted.
- All formatted the sequence files are combined into one file for querying with BLAST.

Step 2. BLAST searches.

Files in FASTA format (from **step 1**) are used as BLAST queries against the full genome sequence of strain Mu50 as follows:

(a) The genome sequence file (NC_002758.fna) is downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Staphylococcus_aureus_Mu50/).

(b) The genome sequence files are formatted using the BLAST program *formatdb*. The command line arguments used are
- "formatdb –i –p F NC_002758.fna"
- where –i Input file(s) for formatting (this parameter must be set) [File In]
- -p Type of file (T—protein or F—nucleotide)
- The resultant files produced are NC_002758.fna.nhr, NC_002758.fna.nin, and NC_002758.fna.nsq

(c) BLAST (blastn) is performed using the formatted DNA sequences produced in **step 1** and the *Staphylococcus aureus* Mu50 genome sequence as its database (subject):
- "blastall -p blastn –m 8 -d NC_002758.fna –i query_transposon_sequences.txt -o query_transposon_sequences.out"
  - Where –p Program Name
  - -d Database
  - -i Input file
- -o Output file
  - -m 8 Output format in tabular form

Step 3. Parsing BLAST output.

The raw BLAST output is parsed to find significant similarity with a query sequence. The BLAST output file presents information in tabular format and for each sequence lists the query id, subject id, % identity, alignment length, mismatches, gap openings, query start, query end, subject start, subject end, e value, and bit score. In the BLAST output file, query_transposon_sequences.out, only hits with an E-value less than or equal to 1e-05 are considered significant and placed into file (query_transposon_sequences.out.significant). Query sequences with single or multiple hits in the genome are listed separately in this output file.

Step 4. Mapping BLAST hit locations onto the genome (**Note 6**).

The file NC_002758.ptt is downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Staphylococcus_aureus_Mu50/). This file contains the following information for each predicted ORF: location in the Mu50 chromosome, name, locus name, strand information, the corresponding protein ID, function, and length. This file can be used to determine positions of transposon insertion in the genome using BLAST hit locations from **step 3**. For each BLAST hit location (subject start and subject end), the positions are searched for in the .ptt file. If these positions overlap:
- only one gene, then that gene is reported and noted as such,
- multiple genes, then those genes are reported and noted as such,
- one or more genes and intergenic region(s), then that gene(s) and the gene(s) neighboring that intergenic region(s) are reported and noted as such.

Notes

1. Stability of strains carrying both plasmids and frequency of transposition: Multiple Newman transformants carrying plasmids pBursa and pFA545 are grown and frozen at –80°C. These

strains can be streaked on agar plates at 30°C. Whereas transposition frequency is ~10^{-6} at 43°C, it is significantly lower at 30°C (less than 10^{-12}).

2. General problems with transposition procedure. Typically, after growing cells harboring pBursa and pFA545 at 43°C and plating on $\text{TSA}_{\text{erm10}}$, about 50 colonies should be observed. However, pBursa can integrate into the chromosome at the site of the *pre* gene (SAV0031 in Mu50) that encodes plasmid recombinase. Integration occurs at all temperatures (30°C to 43°C), causing a larger number of colonies to appear at 43°C after plating candidate transposants on $\text{TSA}_{\text{erm10}}$. It is advisable to regularly sample candidate transposants for the loss of chloramphenicol resistance (plasmids integrating at *pre* site do not lose the chloramphenicol resistance marker). Following **step 12** (**Section 3.2**), isolated colonies can be streaked in parallel on $\text{TSA}_{\text{erm10}}$ and $\text{TSA}_{\text{erm10chl2.5}}$ and incubated at 43°C overnight. More than 98% of all candidates should grow on $\text{TSA}_{\text{erm10}}$ plates and fail to grow on $\text{TSA}_{\text{erm10chl2.5}}$ plates.
3. The size of library and essential genes: It is estimated that 25,000 to 30,000 isolates should represent a complete library for a genome encompassing ~2600 genes (**Chapter 22**). This task can be accomplished within 6 months if the mutants are grown and stored individually and within 1 month if the mutants are pooled. Strains can be stored in 96-well plates with freezing solution at −80°C. Comparison of the unfinished *bursa aurealis* Newman library ***(12)*** with the reported characterization of essential genes in *Bacillus subtilis* ***(31)*** identified an overlap of about 150 to 200 homologous genes.
4. Stability of mutations and second site suppressors: Transposon insertions generated using *bursa aurealis* are very stable and do not undergo secondary transposition events. Isolates with more than one stable transposon insertion are rare. Sequence analysis suggests that mutants with two transposon insertions represent less than 1% of the mutant population. However, second site suppressors often occur as a result of decreased fitness caused by disruption of some genes by the transposon. Often, such mutations cause cells to exhibit temperature-sensitive phenotypes. Because cells are grown at 43°C for long periods of time, this is hardly avoidable. We recommend transduction of alleles of interest (in particular, those that are used to assay virulence in animal models of infection) into original *S. aureus* Newman using bacteriophage Φ85 ***(12)***.
5. Use of *bursa aurealis* in other Gram-positive microorganisms: Technically, plasmids pBursa and pFA545 can be used to transform other Gram-positive bacteria. We have recently shown that the described procedure can be utilized for transposon mutagenesis is *Bacillus anthracis* strain Sterne ***(32)***.
6. Sequence analysis: Because the genome of strain Newman (and many other laboratory strains and clinical isolates) have not yet been sequenced, other staphylococcal genomes currently available in GeneBank have to be used for mapping transposition sites. Staphylococcal genomes differ, especially in pathogenicity islands, prophages, and resistance cassettes. In case of strain Newman, about 1% of sequenced insertion sites do not reveal homology to published genome sequences, suggesting the presence of genes specific to Newman.

Acknowledgments

The authors thank Andrea Dedent for careful reading of the manuscript. E.M.G. acknowledges support from the U.S. Department of Energy under contract W-31-109-ENG-38. O.S. and D.M. acknowledge support from University of Chicago Seed Project Awards.

References

1. Fitzgerald, J. R., Sturdevant, D. E., Mackie, S. M., Gill, S. R., and Musser, J. M. (2001) Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 8821–8826.
2. Baba, T., Takeuchi, F., Kuroda, M., Yuzawa, H., Aoki, K., Oguchi, A., et al. (2002) Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet* **359**, 1819–1827.
3. Kuroda, M., Ohta, T., Uchiyama, I., Baba, T., Yuzawa, H., Kobayashi, L., et al. (2001) Whole genome sequencing of methicillin-resistant *Staphylococcus aureus*. *Lancet* **357**, 1225–1240.
4. Archer, G. L. (1998) *Staphylococcus aureus*: a well-armed pathogen. *Clin. Infect. Dis.* **26**, 1179–1181.
5. Ní Eidhin, D., Perkins, S., Francois, P., Vaudaux, P., Höök, M., and Foster, T. J. (1998) Clumping factor B (ClfB), a new surface-located fibrinogen-binding adhesin of *Staphylococcus aureus*. *Mol. Microbiol.* **30**, 245–257.
6. Ton-That, H., Mazmanian, S. K., and Schneewind, O. (2001) The role of sortase enzymes in Gram-positive bacteria. *Trends Microbiol.* **9**, 101–102.
7. Ruvkun, G. B., and Ausubel, F. M. (1981) A general method for site-directed mutagenesis in prokaryotes. *Nature* **289**, 85–88.
8. Foster, T. J. (1998) Molecular genetic analysis of staphylococcal virulence. *Methods Microbiol.* **27**, 432–454.
9. Bae, T., and Schneewind, O. (2006) Allelic replacement in *Staphylococcus aureus* with inducible counter-selection. *Plasmid* **55**, 58–63.
10. Schwan, W. R., Coulter, S. N., Ng, E. Y., Lnghorne, M. H., Ritchie, H. D., Brody, L. L., et al. (1998) Identification and characterization of the PutP proline permease that contributes to *in vivo* survival of *Staphylococcus aureus* in animal models. *Infect. Immun.* **66**, 567–572.
11. Mei, J. M., Nourbakhsh, F., Ford, C. W., and Holden, D. W. (1997) Identification of *Staphylococcus aureus* virulence genes in a murine model of bacteraemia using signature-tagged mutagenesis. *Mol. Microbiol.* **26**, 399–407.
12. Bae, T., Banger, A. K., Wallace, A., Glass, E. M., Aslund, F., Schneewind, O., and Missiakas, D. M. (2004) *Staphylococcus aureus* virulence genes identified by *bursa aurealis* mutagenesis and nematode killing. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 12312–12317.
13. Ji, Y., Zhang, B., Van, S. F., Horn, W. P., Woodnutt, G., Burnham, M. K., and Rosenberg, M. (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science* **293**, 2266–2269.
14. Forsyth, R. A., Haselbeck, R. J., Gohlsen, K. L., Yamamoto, R. T., Xu, H., Trawick, J. D., et al. (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol. Microbiol.* **43**, 1387–1400.
15. Robertson, H. M., and Lampe, D. J. (1995) Recent horizontal transfer of a mariner transposable element among and between *Diptera* and *Neuroptera*. *Mol. Biol. Evol.* **12**, 850–862.
16. Lampe, D. J., Churchill, M. E., and Robertson, H. M. (1996) A purified *mariner* transposase is sufficient to mediate transposition *in vitro*. *EMBO J.* **15**, 5470–5479.
17. Trieu-Cuot, P., Poyart-Salmeron, C., Carlier, C., and Courvalin, P. (1990) Nucleotide sequence of the erythromycin resistance gene of the conjugative transposon Tn1545. *Nucl. Acids Res.* **18**, 3660.

18. Fitzgerald, S. N., and Foster, T. J. (2000) Molecular analysis of the *tagF* gene, encoding CDP-glycerol:poly(gycerophosphate) glycero-phosphotransferase of *Staphylococcus epidermidis* ATCC14990. *J. Bacteriol.* **182**, 1046–1052.
19. Villafane, R., Bechhofer, D. H., Narayanan, C. S., and Dubnau, D. (1987) Replication control genes of plasmid pE194. *J. Bacteriol.* **169**, 4822–4829.
20. Iordanescu, S. (1975) Recombinant plasmid obtained from two different, compatible staphylococcal plasmids. *J. Bacteriol.* **124**, 597–601.
21. Peschel, A., Ottenwalder, B., and Gotz, F. (1996) Inducible production and cellular location of the epidermin biosynthetic enzyme EpiB using an improved staphylococcal expression system. *FEMS Microbiol. Lett.* **137**, 279–284.
22. Janzon, L., and Arvidson, S. (1990) The role of the delta-lysin gene (hld) in the regulation of virulence genes by the accessory gene regulator (agr) in *Staphylococcus aureus*. *EMBO J.* **9**, 1391–1399.
23. Kullik, I., Giachino, P., and Fuchs, T. (1998) Deletion of the alternative sigma factor sigmaB in *Staphylococcus aureus* reveals its function as a global regulator of virulence genes. *J. Bacteriol.* **180**, 4814–4820.
24. Melton, D. A., Krieg, P. A., Rebagliati, M. R., Maniatis, T., Zinn, K., and Green, M. R. (1984) Efficient in vitro synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacteriophage SP6 promoter. *Nucleic Acids Res.* **12**, 7035–7056.
25. Novick, R. P., Edelman, I., and Lofdahl, S. (1986) Small *Staphylococcus aureus* plasmids are transduced as linear multimers that are formed and resolved by replicative processes. *J. Mol. Biol.* **192**, 209–220.
26. Iordanescu, S., Surdeanu, M., Della Latta, P., and Novick, R. (1978) Incompatibility and molecular relationships between small Staphylococcal plasmids carrying the same resistance marker. *Plasmid* **1**, 468–479.
27. Duthie, E. S., and Lorenz, L. L. (1952) Staphylococcal coagulase: mode of action and antigenicity. *J. Gen. Microbiol.* **6**, 95–107.
28. Peng, H. L., Novick, R. P., Kreiswirth, B., Kornblum, J., and Schlievert, P. (1988) Cloning, characterization, and sequencing of an accessory gene regulator (*agr*) in *Staphylococcus aureus*. *J. Bacteriol.* **170**, 4365–4372.
29. Waldron, D. E., and Lindsay, J. A. (2006) Sau1: a novel lineage-specific type I restriction-modification system that blocks horizontal gene transfer into *Staphylococcus aureus* and between *S. aureus* isolates of different lineages. *J. Bacteriol.* **188**, 5578–5585.
30. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
31. Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., Arnaud, M., et al. (2003) Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4678–4683.
32. Tam, C., Glass, E. M., Anderson, D. M., and Missiakas, D. (2006) Transposon mutagenesis of *Bacillus anthracis* strain Sterne using *Bursa aurealis*. *Plasmid* **56***,* 74–77.

8

Multipurpose Transposon Insertion Libraries for Large-Scale Analysis of Gene Function in Yeast

Anuj Kumar

Summary

Transposons have long been recognized as useful laboratory tools facilitating genome-scale studies of gene function. Relative to traditional methods, transposon mutagenesis offers a rapid and economical means of generating large numbers of independent insertions in target DNA through minimal experimental manipulation. In particular, the transposon insertion library described here is an excellent tool for the analysis of gene function on a large scale in the budding yeast *Saccharomyces cerevisiae*. The transposon utilized in this library is multifunctional, such that the library can be used to screen for disruption phenotypes while also providing a means to generate epitope-tagged alleles and, in many cases, conditional alleles. Provided here are complete protocols by which the transposon insertion library may be used to screen for mutant phenotypes in yeast as well as accompanying protocols describing a means of identifying transposon insertion sites within strains of interest. In total, this insertion library is a singularly useful tool for genome-wide functional analysis, and the general approach is applicable to other organisms in which transforming DNA tends to integrate by homologous recombination.

Key Words: β-gal assays; conditional mutants; Cre-*lox* recombination; epitope tagging; insertion-based screens; insertion mutagenesis; *Saccharomyces cerevisiae*; shuttle mutagenesis; transposon; transposon tagging; vectorette PCR.

1. Introduction

Gene function can be investigated through a variety of approaches encompassing studies of disruption phenotypes ***(1–5)***, conditional alleles ***(6)***, and protein localization ***(7, 8)***. In particular, essential genes can be identified through large-scale gene disruption analysis ***(9)***, and the functions of these genes can be studied further through the use of conditional mutants ***(10)*** and epitope-tagged alleles ***(8, 11)***. For this purpose, transposon mutagenesis has been used to construct a plasmid-based library of mutant alleles, which could facilitate each of these studies in the budding yeast *Saccharomyces cerevisiae* ***(9, 12)***. The transposon used to generate this library is multifunctional; the features of this transposon are detailed in **Section 3.1**. Briefly, the transposon carries a reporter gene and an epitope tag, enabling the transposon insertion library to be used for each

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

of the studies described above. Application of this library is simple. The insertion library can be introduced into a yeast strain by standard methods of DNA transformation; the transposon-mutagenized yeast genomic DNA integrates into the genome by homologous recombination. If desired, Cre-*lox* recombination can be used to remove most of the transposon insertion, leaving behind a sequence encoding an epitope tag. Resulting yeast strains can be screened for phenotypes and/or protein localization, and the site of transposon insertion within these strains can be identified by PCR or other approaches.

Relative to other methods of genome-wide screening, transposon-insertion libraries offer several advantages. Classic methods of chemical treatment or ultraviolet irradiation yield mutations that are difficult to identify within strains of interest, whereas transposon insertions can be located easily by polymerase chain reaction (PCR) amplification or plasmid rescue *(9, 12, 13)*. Genome-wide collections of gene deletions (*[1, 14]* and **Chapter 14**) and fluorescent protein fusions *(7)* also offer an alternative to traditional screens; however, the transposon insertion library described here is more versatile, providing epitope-tagged alleles, gene disruptions, and conditional mutants in a single mutant collection. Finally, by random transposon mutagenesis, we often recover multiple strains with independent insertions at distinct sites within a single gene. Multiple insertion alleles of a gene can be more informative than a single gene deletion as these multiple alleles can be used to define domains within a protein *(9)*.

This chapter presents protocols for the application of transposon-insertion libraries for genome-wide screens of gene function in *S. cerevisiae*. Although the libraries described here are specific for yeast, the general approach can be adopted for other organisms as well in which transforming DNA tends to integrate by homologous recombination.

2. Materials

1. Tris/EDTA (TE) buffer, sterile: 10 mM Tris-Cl, pH 8.0, 1 mM EDTA, pH 8.0.
2. Luria Broth (LB) medium, sterile: 10 g tryptone, 5 g yeast extract, 5 g NaCl, 1 mL 1 N NaOH, add water to 1 L.
3. Selective plates with antibiotics (e.g., tetracycline, kanamycin) as indicated.
4. *Not* I and *Alu* I restriction endonucleases (New England Biolabs, Ipswich, MA).
5. One-step buffer: 0.2 M lithium acetate, 40% (w/v) PEG 4000, 100 mM 2-mercaptoethanol.
6. DNA, RNA: sonicated salmon sperm DNA, 2 mg/mL; yeast tRNA.
7. Synthetic Complete (SC) dropout medium, sterile: per liter, 1.3 g dropout powder, 1.7 g yeast nitrogen base (BD–Difco, Franklin Lakes, NJ) without amino acids/ammonium sulfate, 5 g ammonium sulfate, 20 g dextrose.
8. Yeast/peptone/dextrose (YPD) medium, sterile: mix 10 g yeast extract and 20 g bactopeptone in 950 mL ddH2O in a 2-L flask. Autoclave and add 50 mL 40% dextrose.
9. Yeast/peptone/adenine/dextrose (YPAD) medium, sterile: 1% yeast extract, 2% peptone, 2% dextrose, 80 mg/L adenine.
10. Whatman 3MM filter paper (Whatman, Clifton, NJ).
11. 9-cm and 15-cm glass Petri dishes.
12. Chloroform.

13. 5-Bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal; 120 μg/mL).
14. X-gal plates, sterile: per liter, 1.7 g yeast nitrogen base without amino acids/ammonium sulfate, 5 g ammonium sulfate, 20 g dextrose, 20 g agar, 0.8 g dropout powder, NaOH pellet. Add water to 900 mL and autoclave. Add 100 mL potassium phosphate, pH 7.0, and 2 mL 20 mg/mL X-gal prepared in 100% *N,N*-dimethylformamide.
15. 5-Fluoro-orotic acid (5-FOA) plates, sterile: bacto-yeast nitrogen base (0.67%), dropout mix–Ura (0.2%), glucose (2%), uracil (50 μg/mL), 5-FOA (0.1%), bacto-agar (2%), water.
16. Clinical tabletop centrifuge.
17. 45°C water bath.
18. 96-well plate reader (optional).
19. Oligonucleotide primers.
20. Thermal cycler.
21. Agarose gel equipment.
22. Sporulation medium: 1% (w/v) potassium acetate, 0.1% bacto-yeast extract, 0.05% glucose, 2% bacto-agar, water to 1 l, with appropriate nutritional supplements (at the level of 25% of those used for standard SC plates) depending upon the auxotrophies of the strain.
23. β-Glucuronidase (Sigma-Aldrich, St. Louis, MO).
24. Oligo ABP1: GAAGGAGAGGACGCTGTCTGTCGAAGGTAAGGAACGGACGA-GA GAAGGGAGAG; Oligo ABP2: GACTCTCCCTTCTCGAATCGTAACCG-TTCGTAC GAGAATCGCTGTCCTCTCCTTC.
25. Universal Vectorette (UV) Oligo: CGAATCGTAACCGTTCGTACGAGAATCGCT.

3. Methods

The methods presented here will describe (1) the design and application of multipurpose bacterial transposons for mutagenesis of yeast DNA, (2) the introduction of transposon insertion libraries into yeast for subsequent functional analysis, and (3) an example phenotypic screen using these mutagenized yeast strains.

3.1. Multipurpose Bacterial Transposons

The development of a multifunctional transposon insertion library is described in **Section 3.1.1** to **Section 3.1.2**. This includes (a) the description of the donor transposon construct and (b) a brief description of the *in vitro* mutagenesis protocol used to generate the yeast insertion library.

3.1.1. Transposon Design

Multipurpose transposons for mutagenesis of yeast DNA have been constructed from the bacterial transposons Tn*3* and Tn7. Tn*3* mutagenesis is performed *in vivo* in *Escherichia coli* ***(9)***; the Tn*7* system has been adapted for use *in vitro* by Nancy Craig's group at Johns Hopkins University ***(15, 16)***. Both transposons have been engineered to carry identical components, and both have been used successfully to generate insertional libraries of yeast genomic DNA; however, statistical analysis of Tn*3* and Tn*7* insertion sites suggests that Tn*7* possesses a less-pronounced bias in target site selection than does Tn3 ***(12)***. Thus, a Tn*7*-based insertional library may provide better genome coverage than a Tn*3*-based library. In particular, this chapter presents protocols for the use of a Tn*7*-based library.

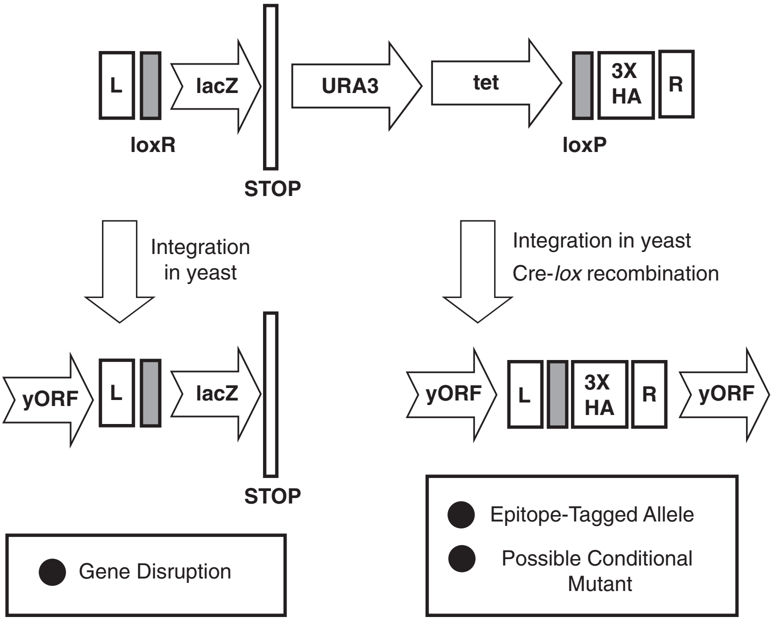


Fig. 1. Schematic diagram of the Tn*7*-derived multipurpose transposon and its applications. The transposon-encoded *lacZ* reporter is indicated as an arrow, as are the bacterial and yeast selectable markers *tet* and *URA3*. The *lacZ* reporter is terminated by a series of stop codons, enabling the construct to be used to generate gene disruptions. Upon Cre-*lox* recombination, the residual transposon construct can be used to generate epitope-tagged alleles and potential conditional alleles. L, Tn*7* left terminus; R, Tn*7* right terminus; 3XHA, sequence encoding three copies of the HA epitope.

The Tn*7*-derived mini-transposon (mTn) is multifunctional in that it can be used to generate a variety of mutant alleles from a single insertion (**Fig. 1**). The transposon itself is bounded by Tn*7* terminal sequences that act as substrates and binding sites for recombination proteins mediating Tn*7* transposition. For selection in *Escherichia coli* and yeast, this transposon carries the *tet* and *URA3* genes, respectively. The Tn*7* transposon contains a modified form of the reporter gene *lacZ* lacking both its start codon and upstream promoter. Transposon insertion, therefore, may be used to generate a *lacZ* gene fusion and β-galactosidase (β-gal) activity, provided that the insertion is oriented such that the *lacZ* reporter is in frame with its surrounding gene. By assaying for β-galactosidase activity, this Tn*7* transposon may serve as a gene trap identifying transcribed and translated sequences within the yeast genome. The *lacZ* reporter is terminated by a series of stop codons such that mTn insertion creates a gene truncation. In addition, the Tn7 transposon contains two *lox* elements, one located near each mTn end; one *lox* site is also internal to sequence encoding three copies of an epitope from the influenza virus hemagglutinin protein (the HA epitope). As *lox* sites are target sequences for the site-specific recombinase Cre, Cre-*lox* recombination may be used in yeast to reduce the full-length 6-kb transposon to a small 99-codon read-through insertion encoding three copies of the HA epitope (the HAT tag). In this manner, mTn-mediated disruption alleles may be converted in yeast to epitope-tagged alleles and, potentially, conditional mutations.

3.1.2. Transposon Insertion Library

This Tn*7*-derived mini-transposon transposon has been used to mutagenize a plasmid-based library of yeast genomic DNA by protocols outlined in Kumar et al. (*12*). In brief, nonspecific Tn*7* transposition is achieved *in vitro* using three purified proteins: TnsA, TnsB, and a TnsC gain-of-function mutant. Paired with the TnsAB transposase and appropriate cofactors (i.e., ATP and Mg^{2+}), the TnsC mutant permits nondirected transposition. In addition, the transposon is subject to transposition immunity, wherein DNA molecules containing at least one transposon terminus are immune from further insertions. Collectively, these properties of the *in vitro* Tn*7* system can be exploited to generate a library of plasmids or PCR products, each bearing a single transposon insertion.

Tn*7* transposon mutagenesis of yeast genomic DNA is performed as follows:

1. Approximately 400 ng of a plasmid-based library of yeast genomic DNA (Kan^r) is mixed with 25 ng of a gain-of-function TnsC mutant ($TnsC^{A225V}$) in a 100-μL reaction volume containing the following (at final concentration): 26 mM HEPES, 4.2 mM Tris pH 7.6, 50 μg/mL BSA, 2 mM ATP, 2.1 mM DTT, 0.05 mM EDTA, 0.2 mM $MgCl_2$, 0.2 mM CHAPS, 28 mM NaCl, 21 mM KCl, and 1.35% glycerol.
2. This mixture is "preincubated" at 30°C for 20 min.
3. Subsequently, 40 ng TnsA, 25 ng TnsB, 15 mM MgAc, and 100 ng donor plasmid (Tet^r) are added to this mixture and incubated at 30°C for an additional 2 h.
4. Products are phenol extracted and ethanol precipitated in the presence of 5 μg yeast tRNA.
5. Precipitated product is then collected, washed, dried, and resuspended in 20 μL water with RNAseA.
6. This reaction is typically repeated five or six times per pool of library DNA; the resulting product is introduced by electroporation into competent *E. coli*.
7. Transformants are plated on LB medium supplemented with tetracycline (3 μg/mL) and kanamycin (40 μg/mL).
8. Transformants are scraped into a cell suspension and stored as frozen stock in 15% glycerol.

The mutagenesis protocol described above was applied individually to 10 pools of a plasmid-based yeast genomic library derived from a strain lacking both its mitochondrial genome [r^-] as well as 2-μm DNA [cir^0]. Each mutagenized pool contains 5 genome equivalents of genomic DNA; thus, the library in total consists of 50 genome equivalents encompassing in excess of 300,000 independent insertions.

The library can be obtained from the author free of charge upon request.

3.2. Generating Yeast Mutants from the Transposon Insertion Library

In order to utilize the transposon insertion library for functional analysis, the insertion alleles must be introduced into yeast by standard methods of DNA transformation. Subsequently, the yeast transformants may be screened to identify potential gene disruptions and insertions in-frame with host genes (**Note 1**). Strains carrying in-frame

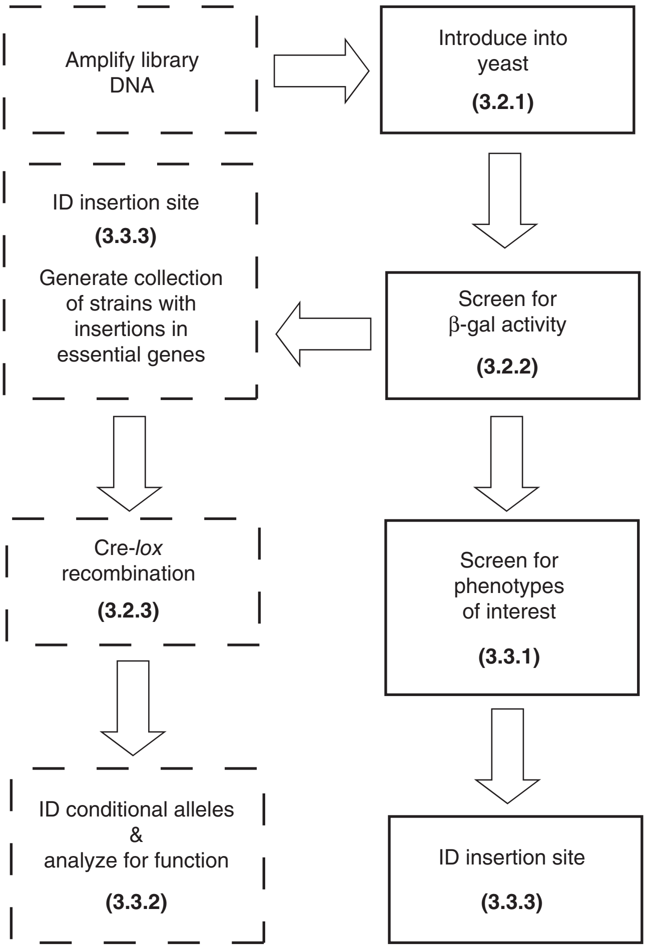

Amplify library DNA
Introduce into yeast
(3.2.1)
ID insertion site
(3.3.3)
Generate collection of strains with insertions in essential genes
Screen for β-gal activity
(3.2.2)
Cre-*lox* recombination
(3.2.3)
Screen for phenotypes of interest
(3.3.1)
ID conditional alleles & analyze for function
(3.3.2)
ID insertion site
(3.3.3)

Fig. 2. Outline of the steps by which the transposon insertion library can be used to screen for phenotypes of interest and/or generate epitope-tagged alleles and conditional alleles for the analysis of essential genes. Optional steps are indicated inside of dashed lines. Chapter sections corresponding with each protocol are shown in boldface.

insertions can be used to generate epitope-tagged alleles by Cre-*lox*–mediated modification of the integrated transposon. A diagrammatic outline of these steps is provided in **Figure 2**. These protocols are described in **Section 3.2.1** to **Section 3.2.3**.

3.2.1. Introduction of the Library into Yeast by DNA Transformation

Approximately 1 µg of each pool of library DNA will be supplied to users; thus, it may be necessary to obtain a greater quantity of library DNA for transformation of yeast. If desired, the library DNA may be introduced into any tetracycline- and kanamycin-sensitive *E. coli* strain by standard transformation procedures.

1. Select transformants on LB medium supplemented with tetracycline (3 µg/mL) and kanamycin (40 µg/mL) using plates 14 cm in diameter. In total, approximately 100,000 transformants should be obtained following overnight growth at 37°C.

2. Elute transformant colonies as follows: Place 6 mL LB medium onto each plate and scrape cells into a homogenous suspension. Dilute an aliquot of this eluate into LB medium supplemented with tetracycline (3 μg/mL) and kanamycin (40 μg/mL) to yield a culture of nearly saturated cell density. Incubate at 37°C with aeration for 2 to 3 h.
3. Isolate plasmid DNA by standard alkaline lysis.
4. Digest approximately 1 μg plasmid DNA with *Not* I. Subsequently, analyze a portion of the reaction mixture by agarose gel electrophoresis to ensure release of mTn-mutagenized yeast DNA from the plasmid vector (**Note 2**). Store the remaining reaction mixture for later use in **step 7**.
5. Grow a 10-mL culture of any desired *ura3* yeast strain to mid-log phase (a density of 10^7 cells/mL or OD_{600} of approximately 1) maintaining appropriate selection if applicable (**Note 3**).
6. Pellet cells in a clinical tabletop centrifuge at 1100 × *g* for 5 min. Wash once with 5 volumes of one-step buffer.
7. Resuspend cells in 1 mL one-step buffer supplemented with 1 mg denatured salmon sperm DNA. Add 100-μL aliquots from this suspension to 0.1 to 1 μg *Not* I–digested plasmid DNA from **step 4** (**Note 4**). Vortex, and incubate at 45°C for 30 min.
8. Pellet cells and subsequently suspend in 400 mL SC-Ura medium. Spread 200-μL aliquots onto SC-Ura plates and incubate at 30°C for 3 to 4 days. Up to 1000 transformants may be recovered per microgram of transforming DNA (**Note 5**).

3.2.2. Screening Yeast Mutants for In-Frame Insertions

Productive transposon insertions within protein coding sequences can be detected by virtue of the *lacZ* reporter encoded within the transposon. Yeast strains containing an insertion in-frame with the surrounding gene will produce *lacZ*-encoded β-galactosidase, provided that the gene is expressed under the given growth conditions. β-Gal activity can be assayed easily as described below.

1. To maximize detection of *lacZ* fusions expressed at low levels, patch transformant colonies onto YPD plates (supplemented with 80 μg/mL adenine if using an *ade2* host strain) at a density of up to 100 colonies per plate.
2. Place a sterile disk of Whatman 3MM filter paper onto a plate of SC-Ura medium; repeat for as many plates as needed. Replicate transformant cells onto filter-covered plates and incubate overnight at 30°C (**Note 6**).
3. Following overnight growth, lift filters from plates and place in the lid of a 9-cm glass Petri dish. Place this lid inside a closed 15-cm Petri dish containing chloroform. Incubate for 10 to 30 min.
4. Place filters colony-side up onto fresh X-gal plates (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside [X-gal; 120 μg/mL], 0.1 M phosphate buffer pH 7.0, 1 mM $MgSO_4$ in 1.6% [w/v] agar). Incubate inverted at 30°C for up to 3 days. After several days of growth, β-gal levels can be reliably estimated from the observed intensity of blue staining (**Note 7**).

3.2.3. Generating Epitope-Tagged and Conditional Alleles by Cre-lox Recombination

Strains bearing an in-frame transposon insertion may be used to derive corresponding strains with epitope-tagged proteins by Cre-*lox* recombination in yeast. The phage P1 Cre recombinase can be expressed exogenously from plasmid pGAL-*cre* (available from the author); on this plasmid, *cre* is under transcriptional control of the *GAL* promoter, so that galactose induction may be used to drive *cre* expression. Following induction on galactose, cells having undergone Cre-mediated recombination (and loss of the mTn-encoded *URA3* marker) may be selected on medium containing 5-fluoro-orotic acid (5-FOA). The residual transposon insertion may also effectively generate hypomorphic or conditional mutants. This procedure is described below.

1. Transform the mTn-mutagenized *ura3 leu2* host strain with pGAL-*cre* (*amp*, *ori*, *CEN*, *LEU2*); subsequently, select transformants on SC-Leu-Ura dropout medium.
2. To derepress the *GAL* promoter, inoculate transformants into 2 mL SC-Leu-Ura medium with 2% raffinose as its carbon source. Incubate at 30°C with aeration until the culture has grown to saturation.
3. Dilute cultures 100-fold into SC-Leu medium with 2% galactose as its carbon source. As a control, dilute an aliquot of the same culture 100-fold into 2 mL SC-Leu medium with 2% glucose as its carbon source. Grow cultures for 2 days at 30°C with aeration.
4. If visible growth is apparent, dilute cultures 100-fold in sterile water and withdraw a 10-μL aliquot. If no growth is apparent, withdraw a 10-μL aliquot from the undiluted culture. Spot onto a 5-FOA plate, and isolate single colonies by streaking the droplet. Dilute cultures grown in 2% glucose 100-fold in sterile water, withdraw a 10-μL aliquot, spot, and streak onto a 5-FOA plate. Incubate 5-FOA plates at 30°C until growth is visible on those plates inoculated with strains grown in galactose (**Note 8**).
5. Single colonies from strains having lost the mTn-encoded *URA3* marker (exclusively following galactose induction) may be saved as stock in 15% glycerol at −70°C.

3.3. Phenotypic Analysis of Yeast Insertion Mutants

A general protocol for screening yeast insertion mutants is presented in **Section 3.3.1** to **Section 3.3.2**. This includes (a) considerations in planning a phenotypic study of gene disruptions, (b) protocols for identifying and studying conditional mutants, and (c) protocols for the identification of the transposon insertion site within strains of interest.

3.3.1. Analysis of Disruption Phenotypes

Yeast insertion mutants can be screened for any number of desired phenotypes by growing the strains under appropriate conditions. As the specifics of this screen will necessarily vary depending upon the chosen growth conditions, a general outline is provided here.

Yeast mutants may be conveniently assayed in arrayed format if desired. Liquid cultures of yeast transformants can be grown in 96-well microtiter plates to mid-log phase under environmental stress; for example, resistance/sensitivity to a drug may be determined by growing cells in media supplemented with that drug. Identical cultures can be grown in parallel in 96-well format under standard growth conditions (e.g., in normal growth media lacking the environmental stress). To obtain a quantitative measure of strain fitness, cell density can be measured (at OD_{600}) using a 96-well plate reader. The cell density of treated versus mock-treated samples provides an indication of strain sensitivity or resistance. This very simple screening approach can be modified as needed.

3.3.2. Analysis of Conditional Mutants

The yeast insertion library can also be used to investigate cell functions associated with essential genes *(17)*. If a diploid strain of yeast is chosen for transformation, insertions in essential genes can be recovered, provided the heterozygous mutant is viable. By Cre-*lox* recombination, the full-length transposon can be modified into an epitope-insertion element, with the added benefit of generating conditional mutants in some cases (**Note 9**). Conditional mutants may be identified and studied as follows.

1. Heterozygous diploid strains carrying an epitope-insertion element in an essential gene should be sporulated to determine whether the insertion disrupts gene function. Briefly, inoculate cells from the strain of interest in 1.5 mL sporulation media. Incubate cultures on a roller drum for 3 to 5 days at room temperature (**Note 10**).
2. Harvest sporulated cultures (3000 ×*g* for 5 min) and wash. Resuspend tetrads in 1 mL sterile water.
3. Digest a 100-μL aliquot of resuspended tetrads in 5 μL of β-glucuronidase (134.6 units/mL) at room temperature for 15 to 20 min.
4. After incubation, spread 8 μL of the reaction mixture onto an appropriate plate. Dissect tetrads and incubate plates at room temperature for 2 to 3 days.
5. Score haploid segregants for viability. For example, if using the lab strain BY4741, score segregation of the *met*, *lys*, and *MAT* loci by replica plating onto SC-Met, SC-Lys, and SC-His plates spread with lawns of mating-type testers. Identify strains exhibiting $4^+:0^-$ segregation of viability.
6. Select for further analysis four haploids derived from one tetrad, as well as strains that are ***MATamet15*** and *MATalys2*. The complete tetrad serves as a control, confirming that the epitope-insertion element is segregating $2^+:2^-$ as expected.
7. For each strain analyzed, use PCR (or other methods) to verify presence of the insertion element and loss of the wild-type allele in appropriate haploids.
8. Test viable haploid progeny carrying the epitope-insertion element for the desired hypomorphic or conditional phenotype, such as temperature sensitivity.

3.3.3. Identifying the Transposon Insertion Site in a Strain of Interest

Several approaches may be used to identify the precise site of transposon insertion within a strain of interest (e.g., one exhibiting a desired phenotype from the screens described above). For example, insertion sites may be identified through direct genomic sequencing of mTn-mutagenized strains using a transposon-specific primer *(18)*. Alternatively, PCR-based methods, such as the vectorette approach *(19, 20)*, can be utilized to identify transposon insertion sites.

In vectorette PCR (**Fig. 3**), genomic DNA is digested with a blunt-end restriction endonuclease possessing a 4- to 6-base-pair recognition sequence. Blunt-ended DNA fragments are ligated to a pair of annealed primers containing a nonhomologous central region; these primer pairs form “anchor bubbles” flanking each genomic fragment. PCR is then performed using a primer complementary to the transposon and a primer identical to the sequence within the anchor bubble. During the initial round of amplification, only the mTn primer can bind its template; however, during subsequent cycles, the anchor bubble primer can anneal to the extended mTn primer, resulting in selective amplification of DNA sequence adjacent to the point of transposon insertion.

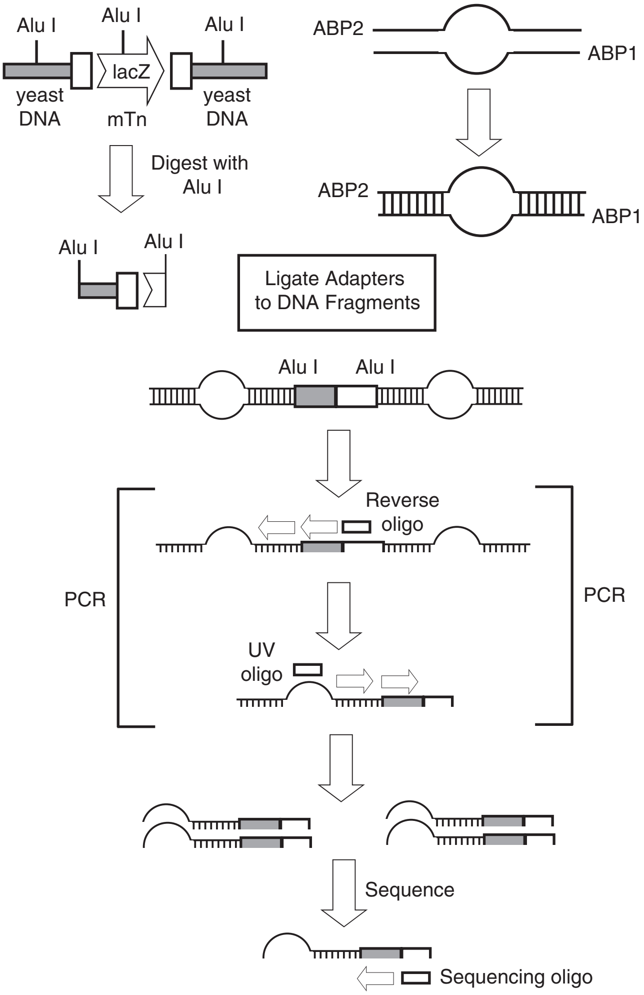


Fig. 3. Identification of transposon insertion sites by vectorette PCR. Here, *Alu* I is used to cleave the yeast genomic DNA; however, other endonucleases may be used. Sequences for the indicated oligonucleotides are included in the **Materials** section. Note that many oligonucleotide sequences may be used to reverse prime amplification from the transposon sequence; similarly, many oligonucleotides may be used to sequence the final PCR product. Thus, these oligonucleotide sequences are not indicated in the **Materials** section.

The vectorette PCR protocol provided below should yield approximately 200 to 400 ng of product, constituting sufficient template for 2 to 3 sequencing reactions.

1. Prepare genomic DNA by any standard protocol (**Note 11**). Digest 5 µg of yeast genomic DNA with a blunt-end restriction endonuclease (such as *Alu* I) in a total volume of 20 µL. After overnight digestion, the enzyme is heat-inactivated by incubating 20 min at 65°C.
2. Anneal primers ABP1 and ABP2 to form the adapter anchors by mixing 1 pmol of each primer in 200 µL of annealing buffer containing 10 mM Tris, 10 mM $MgCl_2$, and 50 mM NaCl. Heat the primer mixture for 5 min at 95°C and cool slowly to 37°C.

3. Ligate adapters to the DNA fragments by adding 1 µL of the annealed primers, 0.25 µL of 10 mM ATP, 3 µL of 10× restriction buffer used in the digest, and 24.25 µL H_2O to the 20-µL restriction digest mixture from **step 1**. Incubate the ligation reaction overnight at 16°C.
4. Perform a standard 100-µL PCR reaction using 5 µL from the ligation mixture, 2.5 µL each of the UV primer and a reverse primer complementary to the transposon at 20 µM, 5 U of thermostable polymerase, and 1 µL of deoxynucleotide triphosphates (dNTPs) (at 20 mM each dNTP) in a final volume of 100 µL. The PCR program consists of one cycle of 2 min at 92°C, followed by 35 cycles of 20 s at 92°C, 30 s at 67°C, and 45 to 180 s at 72°C with a final extension of 90 s at 72°C.
5. Analyze PCR products by gel electrophoresis. Extract and purify each PCR product from the agarose gel into a final volume of 30 µL TE. Ten microliters of the purified product is sufficient for one sequencing reaction.

Notes

1. Essential genes can be screened for haploinsufficiency by introducing the insertion library into a diploid strain of yeast. After Cre-*lox* recombination in yeast, the modified epitope-insertion element can be used to generate conditional mutants in some cases. Thus, essential genes can also be studied in this manner using the transposon insertion library.
2. Upon *Not* I digestion and electrophoresis, a distinct 2.1-kb band (corresponding with the vector) and broad 8-kb band should be visible: the broad 8-kb band consists of 2- to 3-kb inserts of yeast genomic DNA carrying the 6-kb mTn construct.
3. Ideally, choose a diploid yeast strain to screen for desired patterns of gene expression. To screen for disruption phenotypes, a haploid strain is often used; from previous studies (*17*), we estimate that 10% of transposon insertions in essential genes are viable. For the eventual analysis of epitope-tagged proteins, choose a *ura3 leu2* strain, as the pGAL-Cre vector carries the *LEU2* gene.
4. Use a small quantity of transforming DNA in order to minimize the generation of transformants containing more than one insertion.
5. To ensure 95% coverage of the genome (without regard to in-frame reporter activity), screen 30,000 to 50,000 colonies. To identify in-frame insertions within at least 95% of all yeast genes, screen approximately 180,000 to 200,000 transformants for β-gal activity.
6. Alternate growth conditions (e.g., growth on sporulation medium) may be substituted as desired.
7. β-Gal activity is typically observed in 12% to 16% of transformants.
8. From previous experience, galactose induction results in Cre-mediated excision of the *URA3* marker in more than 90% of cells analyzed.
9. From a pilot study of 143 heterozygous diploid strains carrying an in-frame epitope-insertion element in a gene essential for yeast cell growth, 28% of essential proteins carrying an in-frame transposon-encoded epitope-insertion element retain at least partial function. Subsequent screening will be necessary to determine whether these epitope-insertion elements have actually generated conditional mutants. Application of this approach to generate conditional mutants in essential genes should be initiated with this understanding.
10. Routinely, cultures from two independent transformant colonies are used; a culture from a third colony may be preserved for use later in resolving any conflicting results.
11. Take care to obtain high-quality DNA, as this is critical to successful PCR amplification.

Acknowledgments

The author thanks Dr. Michael Snyder, in whose laboratory these protocols were developed. This work was supported by awards from the March of Dimes (FY05-1224) and from the American Cancer Society (RSG-06-179-01-MBC) to A.K.

References

1. Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., et al. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906.
2. Spradling, A. C., Stern, D. M., Kiss, I., Roote, J., Laverty, T., and Rubin, G. M. (1995) Gene disruptions using *P* transposable elements: an integral component of the *Drosophila* genome project. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 10824–10830.
3. Martienssen, R. A. (1998) Functional genomics: probing plant gene function and expression with transposons. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 2021–2026.
4. Alonso, J. M., Stepanova, A. N., Leisse, T. J., Kim, C. J., Chen, H., Shinn, P., et al. (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301**, 653–657.
5. Kamath, R., Fraser, A. G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., et al. (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237.
6. Mnaimneh, S., Davierwala, A. P., Haynes, J., Moffat, J., Peng, W. T., Zhang, W., et al. (2004) Exploration of essential gene functions via titratable promoter alleles. *Cell* **118**, 31–44.
7. Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., and O'Shea, E. K. (2003) Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691.
8. Kumar, A., Agarwal, S., Heyman, J. A., Matson, S., Heidtman, M., Piccirillo, S., et al. (2002) Subcellular localization of the yeast proteome. *Genes Dev.* **16**, 707–719.
9. Ross-Macdonald, P., Coelho, P. S., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., et al. (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413–418.
10. Davierwala, A. P., Haynes, J., Li, Z., Brost, R. L., Robinson, M. D., Yu, L., et al. (2005) The synthetic genetic interaction spectrum of essential genes. *Nat. Genet.* **37**, 1147–1152.
11. Hazbun, T. R., Malmstrom, L., Anderson, S., Graczyk, B. J., Fox, B., Riffle, M., et al. (2003) Assigning function to yeast proteins by integration of technologies. *Mol. Cell* **12**, 1353–1365.
12. Kumar, A., Seringhaus, M., Biery, M., Sarnovsky, R. J., Umansky, L., Piccirillo, S., et al. (2004) Large-scale mutagenesis of the yeast genome using a Tn7-derived multipurpose transposon. *Genome Res.* **14**, 1975–1986.
13. Mosch, H. U., and Fink, G. R. (1997) Dissection of filamentous growth by transposon mutagenesis in *Saccharomyces cerevisiae*. *Genetics* **145**, 671–684.
14. Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391.
15. Biery, M., Stewart, F., Stellwagen, A., Raleigh, E., and Craig, N. (2000) A simple *in vitro* Tn7-based transposition system with low target site selectivity for genome and gene analysis. *Nucleic Acids Res.* **28**, 1067–1077.
16. Bachman, N., Biery, M., Boeke, J., and Craig, N. (2002) Tn7-mediated mutagenesis of *Saccharomyces cerevisiae* genomic DNA *in vitro*. *Methods Enzymol.* **350**, 230–247.

17. Kumar, A., des Etages, S. A., Coelho, P., Roeder, G., and Snyder, M. (2000) High-throughput methods for the large-scale analysis of gene function by transposon tagging. *Methods Enzymol.* **328**, 550–574.
18. Horecka, J., and Jigami, Y. (2000) Identifying tagged transposon insertion sites in yeast by direct genomic sequencing. *Yeast* **16**, 967–970.
19. Riley, J., Butler, R., Ogilvie, D., Finniear, R., Jenner, D., Powell, S., et al. (1990) A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. *Nucleic Acids Res.* **18**, 2887–2890.
20. Kumar, A., Vidana, S., and Snyder, M. (2002) Insertional mutagenesis: transposon-insertion libraries as mutagens in yeast. *Methods Enzymol.* **350**, 219–229.

9

How to Make a Defined Near-Saturation Mutant Library. Case 1: *Pseudomonas aeruginosa* PAO1

Michael A. Jacobs

Summary

We have constructed a near-saturation level mutant library for *Pseudomonas aeruginosa* strain PAO1 using Tn*5*-derived transposons mapped to the PAO1 reference sequence. This chapter describes the high-throughput techniques used to generate and map the mutant strains. In addition, an analysis of the utility of this collection is presented based on changes to the annotation for the PAO1 genome in the past years, as well as the citation record for this collection. It is clear that many avenues of research have been accelerated by this collection and that additional large mutant strain collections will further aid in defining gene function and biological processes in pathogens.

Key Words: high-throughput; mutant library; PAO1; *Pseudomonas*; transposon.

1. Introduction

1.1. Utility of Defined Mutant Collections

One of the many exciting opportunities in the genomics era is to use the newly available genome sequence and bioinformatics tools to vastly accelerate the pace of gene function discovery. In microbial research, the standard for determining gene function remains mutant analysis. Two main paradigms have defined this field: forward and reverse genetics. Both methods have contributed immensely to our understanding of gene function in the microbial world. For *Pseudomonas aeruginosa*, the current annotation (as of May 12, 2006) contains 5570 open reading frames (ORFs), and only 42% remain classified as "Hypothetical, unclassified, unknown" (*Pseudomonas aeruginosa* Community Annotation Project).

There is clearly a long way to go, and progress remains hard-won, but the pace of discovery has been increasing greatly in the past 2 years. In the time between the original annotation table (September 1, 2000) and the current table (June 19, 2007), the number of unclassified genes has decreased from 2381 to 2053. This is a significant improvement since 2004 when the number of unclassified genes was 2370. Much of

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

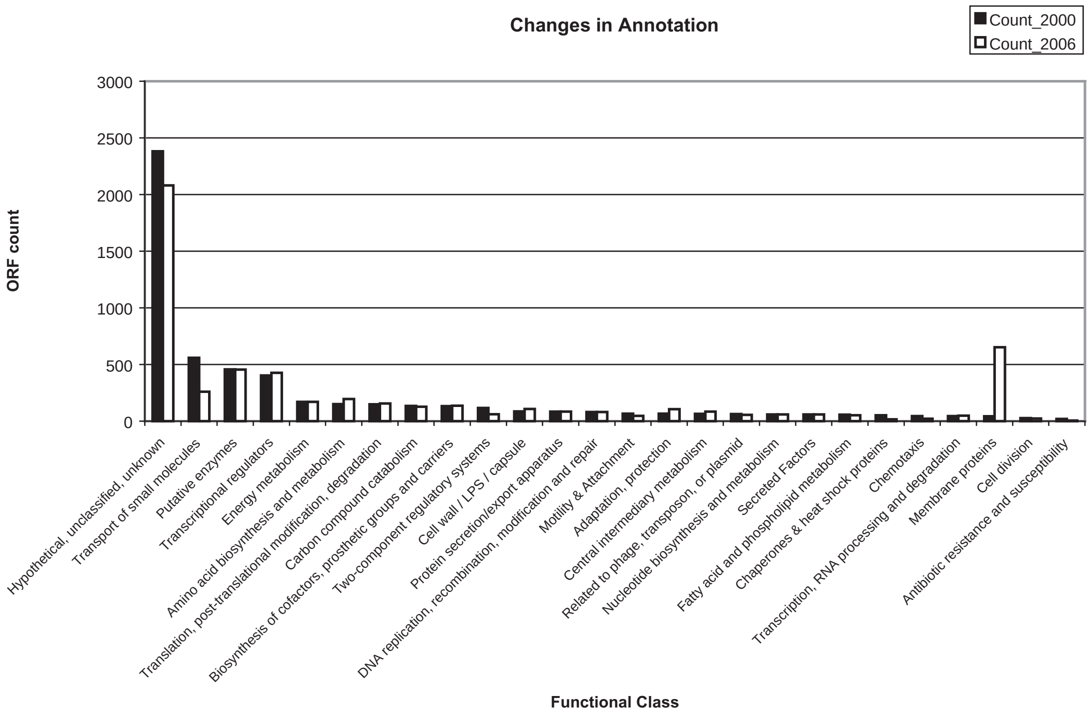


Fig. 1. Changes in annotation. The number of open reading frames (ORFs) in each functional class category is compared between the original published annotation *(1)* and the annotation table as of May 2006, available at the Community Annotation Project managed by the laboratory of Fiona Brinkman (www.pseudomonas.com).

this change may be attributed to a large increase in the number of genes that have been assigned to "membrane proteins" (**Fig. 1**).

We have used the complete genome sequence of *Pseudomonas aeruginosa* PAO1 *(1)* to map a near-saturation collection of defined mutant strains *(2)*. This defined set of strains allows a significant acceleration of discovery using either forward or reverse genetic techniques. Forward genetic approaches of screening the collection are vastly aided by the knowledge about where individual strains contain their transposon insertion. From a reverse genetic standpoint, researchers may quickly obtain mutants from the collection in all genes of interest identified by bioinformatic analysis. It is our hope that distributing the strains to other researchers will help accelerate the research effort overall (see **Notes** section). Since we started shipping mutants until we shut down to transfer the shipping responsibilities (November 2003 through February 2006), we shipped a total of 9236 strains to 217 different researchers in 103 cities in 19 different countries and 28 states within the United States.

1.2. Hurdles to Production

There are a variety of challenges facing researchers who choose to create a large defined mutant collection for a given microbe. The two main categories of challenges are cost and scale-up.

Overall, the cost of sequencing and polymerase chain reaction (PCR) reagents is decreasing, and collaborating with a genome center may significantly reduce the reagent cost of the project. Currently, our total reagent cost per hit is $1.11 at the UW Genome Center (**Table 1**, includes failure rate of 20%). Advances in shrinking reaction sizes and optimizing reagent efficiency with new sequencing machines will likely reduce that cost further. Even so, this is a major price tag for a small lab. On the NCBI Web site (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html), there are nearly 400 completely sequenced microbial genomes with an average size of 3.2 Mbp. If the average gene size is 1000 bp, this leaves the researcher with 3200 genes to knock out. Saturation is approximated at a coverage of 5×, requiring 16,000 insertions, which may cost up to $18,000 per genome. Though large, these prices are not insurmountable, especially given the obvious practical benefit of defined saturated mutant collections (see above). However, these prices are absolutely dependent on the researchers' ability to scale up mutagenesis and mutant mapping in a manner currently practiced mainly at genome centers. Without truly high-throughput technology, the labor and time costs of building these libraries will be unfeasible.

Scaling up the process of creating defined mutants can be divided into two parts: mutant generation/arraying and mapping. Mutagenesis is easily scaleable. Even the most complex mutagenesis protocols usually produce many individual mutants, a property that has enabled phenotype screens using saturated mutagenesis for decades. The main challenge then is to array the individual mutants into high-throughput format for mapping purposes and for long-term storage and individual mutant recovery. For this purpose, a colony-picking robot is a helpful, if not an essential, tool. We used the Qpix robot from Genetix Ltd. (Hampshire, UK). This robot (described in the **Methods** section) arrays thousands of colonies into 384-well plates quickly and efficiently. Arraying colonies by hand is slow, but feasible, depending on time and labor cost constraints. In addition, hand arraying into 384-well format is error-prone compared with the robot. Once colonies are arrayed into plates, they may be easily stored at −80°C until needed for colony PCR, to recover individual mutants, or to replicate the whole plate.

Mapping the mutants to a defined genome location is made feasible by using a finished reference genome. For our *Pseudomonas aeruginosa* genome, the transposon-genome junctions were amplified using semidegenerate nested PCR and sequencing techniques. These techniques were all easily adapted to 384-well format. Sample tracking and data analysis require some bioinformatics expertise. Sample tracking is critical and may be automated or may require that researchers prepare a spreadsheet for each plate to be sequenced. Once the data are generated, the sequence traces (chromats) must be analyzed for quality and then their transposon/genome junctions found. The natural tool for automating this process is a PERL script, which can combine all tasks into one process. Output from the PERL script must be stored in a relational database for further analysis.

The researcher interested in pursuing this activity therefore must have access to high-throughput machinery, a moderate level of computer expertise (or good access to someone who does), and a sufficient budget to carry out the project.

Table 1
Cost of Reagents for High-Throughput Mapping Transposon Mutants to a Reference Genome

Reagent/amount	Cost	Per reaction	Cost/reaction	384 plate (PCR×2)	Per hit
TSG polymerase 600 U	$128.69	0.5 U	$0.11	$82.36	
dNTPs 100 µL 10 mM	$35.00	0.2 µL	$0.07	$53.76	
Primers 2 × 50 nmol	$35.00	5 pmol	$0.00	$1.34	
Total PCR cost	**$198.69**		**$0.18**	**$137.47**	
SAP 1000 U	$110.00	2 U	$0.22	$84.48	
Exonuclease 2500 U	$64.00	10 U	$0.26	$98.30	
Total cleanup cost	$174.00		$0.48	$182.78	
Template prep cost	**$372.69**		**$0.65**	**$320.25**	
BDT 3.1 25,000 Rxns	$79,000.00	1/16th rxn	$0.20	$75.84	
Primer 50 nmol	$17.00	5 pmol	$0.00	$0.65	
Sequencing cost	**$79,017.00**		**$0.20**	**$76.49**	
Current success rate	80%				
Pipetting loss	5%				
Resulting multiplier	1.3				
Inclusive reagent cost per hit					$**1.11**

Reagents comprise a significant portion of the cost of generating a large defined mutant library. PCR, template preparation, and sequencing reagents are shown above, and the approximate cost per hit is calculated given average loss and failure rates of 5% and 20%, respectively.

Table 2
Equipment Needed to Produce a Large Mutant Collection in High-Throughput Scale

Equipment required for high-throughput mutant mapping	Approx. cost
Picking robot: automated arraying of clones into 384-well format	\$200,000–\$500,000
384-well thermocycler	\$5,000–\$20,000
Multichannel pipette	\$1,000
ABI 3730 Autosequencer	\$200,000

2. Materials

1. Critical pieces of high-throughput equipment required are given in **Table 2**.
2. 1× freezer University of Washington Genome Center (UWGC) medium: To a final volume of 4 L, dissolve in de-ionized water, 40 g tryptone-peptone, 20 g yeast extract, 40 g NaCl, 25.2 g K_2HPO_4, 7.2 g KH_2PO_4, 2.0 g sodium citrate, 3.6 g $(NH_4)_2SO_4$, and 176 mL glycerol. Split the 4-L batch into 1-L aliquots, hold in 2-L bottles for autoclave, and autoclave with a 1-h sterilization time. Cool to room temperature before use, add required antibiotics. Chemicals are standard and may be purchased from any vendor, such as Sigma-Aldrich (St. Louis, MO). (*See* **Note 5** for alternate media.)
3. Tetracycline in a final concentration of 20 µg/mL (Sigma-Aldrich).
4. Chloramphenicol in a final concentration of 10 µg/mL (Sigma-Aldrich).
5. 5-Bromo-4-chloro-3-indolyl phosphate (X-phosphate; Sigma-Aldrich).
6. 5-Bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal; Sigma-Aldrich).
7. 384-Well microtiter plate with cover (Genetix Limited).
8. Qrep 384 Pin Replicators (ISC Bio Express, Kaysville, UT).
9. Bioassay dish, case of 20 (Nunc; available from VWR, West Chester, PA).
10. Nalgene Filter: Nalgene catalog no. 126-0045 (available in the United States from VWR).
11. Tsg DNA polymerase (Lamda Biotech, St. Louis, MO).

3. Methods

Many of these methods have been previously summarized in Bailey and Manoil *(3)*, Jacobs et al. *(2)*, and Jacobs and Manoil *(4)*. The main innovation for this project was adapting small-scale mutagenesis mapping protocols to high-throughput equipment. The overall process schematic can be found in **Figure 2**.

3.1. Transposon Mutagenesis

Two Tn*5*-derived transposons were modified to confer tetracycline resistance and included the *phoA* marker gene for identification of exported fusion proteins or the *lacZ* gene for detection of cytoplasmic proteins. In both cases, the transposons make translational fusions when the transposon lands in the correct orientation and frame in an expressed gene. Fusions in *phoA* require that the protein be secreted to be detected, and *lacZ* fusions are detectable primarily when the transposon gene (and resulting fusion) is expressed cytoplasmically. Both transposons contain flanking *loxP* elements to allow conversion into a small 63-codon insertion encoding the influenza hemagglutinin (HA) epitope and a hexahistidine motif. In addition, the transposons contain an outwardly directed promoter element at their 3′ ends to minimize polar effects.

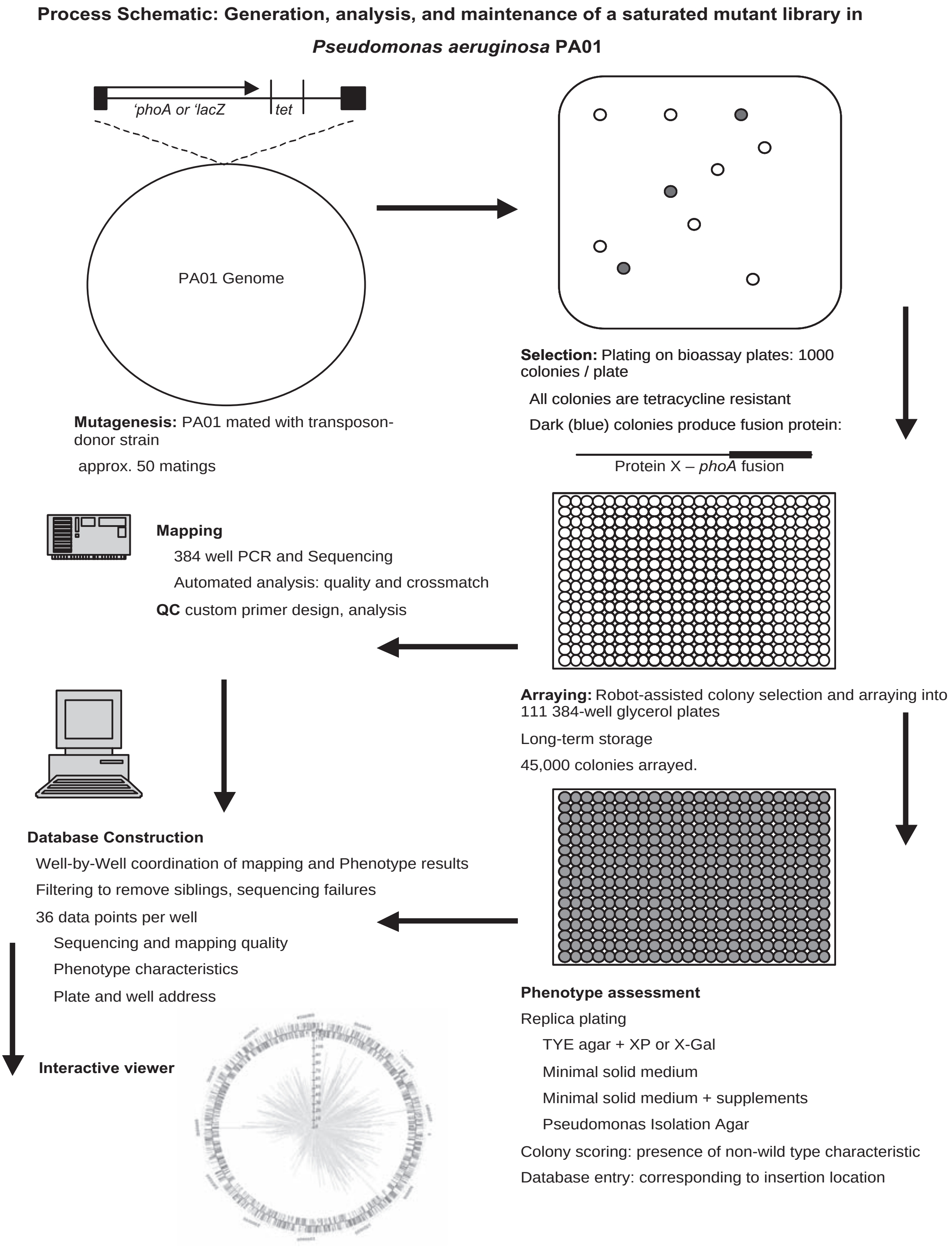

Process Schematic: Generation, analysis, and maintenance of a saturated mutant library in
Pseudomonas aeruginosa PA01
'phoA or 'lacZ
tet
PA01 Genome
Mutagenesis: PA01 mated with transposon-donor strain
approx. 50 matings
Selection: Plating on bioassay plates: 1000 colonies / plate
All colonies are tetracycline resistant
Dark (blue) colonies produce fusion protein:
Protein X – phoA fusion
Mapping
384 well PCR and Sequencing
Automated analysis: quality and crossmatch
QC custom primer design, analysis
Arraying: Robot-assisted colony selection and arraying into 111 384-well glycerol plates
Long-term storage
45,000 colonies arrayed.
Database Construction
Well-by-Well coordination of mapping and Phenotype results
Filtering to remove siblings, sequencing failures
36 data points per well
Sequencing and mapping quality
Phenotype characteristics
Plate and well address
Phenotype assessment
Replica plating
TYE agar + XP or X-Gal
Minimal solid medium
Minimal solid medium + supplements
Pseudomonas Isolation Agar
Colony scoring: presence of non-wild type characteristic
Database entry: corresponding to insertion location
Interactive viewer

Fig. 2. High-throughput mutagenesis and mapping. A process schematic describing the scheme we used to array and map random transposon mutants. Parts of this figure were originally published as Figure 3 in Ref. *4* (p. 125) and are reproduced here with kind permission of Springer Science+Business Media.

Conjugation-generated insertions in the PAO1 chromosome were generated by mating a wild-type PAO1 strain (obtained from B. Iglewski, University of Rochester Medical Center, Rochester, NY) referred to as MPAO1, with strain SM10*pir*/pCM639 (for IS*phoA*/hah insertions) or SM10*pir*/pIT2 (for IS*lacZ* insertions).

The following transposon mutagenesis protocol, refined by Larry Gallagher, Ashley Alwood, and Colin Manoil (University of Washington, Seattle, WA) was used:

1. Grow overnight cultures of MPAO1 in LB (tryptone (10 g/L), yeast extract (5 g/L), NaCl (10 g/L) plus 15 g/L agar) at 42°C without aeration. Grow *Escherichia coli* SM10 *pir*/ pCM639 or pIT2 at 37°C with aeration in LB supplemented with 100 μg/mL ampicillin.
2. The next day, dilute the *E. coli* strain 1 : 10 in fresh LB-amp and grow 45 min at 37°C with aeration.
3. Mix 0.5 mL of the PAO1 and *E. coli* strains and immediately filter through presterilized Nalgene Analytical Test Filter Funnel (0.45-μm nominal pore size). (Do not forget PAO1 alone and *E. coli* alone as controls.) Wash with 1 mL 10 mM $MgSO_4$. Place filter on a prewarmed TYE (or LB agar) plate using sterile forceps.
4. Incubate 1 to 2 h at 37°C.
5. Remove filter with forceps and place in a large sterile test tube containing 1 mL LB. Vortex thoroughly, checking that cell mass is washed into broth.
6. Plate undiluted (as well as 1 : 10 and 1 : 100 diluted cells until the titer is standardized) on (dry) TYE agar containing 10 μg/mL chloramphenicol (to counterselect against the *E. coli* donor strain) and 60 μg/mL tetracycline. When plating on large bioassay plates (Nunc; available from VWR), you may plate the entire mL wash from **step 5** undiluted. For detection of fusion proteins, include the following in your agar plates: X-phosphate (5-bromo-4-chloro-3-indolyl phosphate) for detection of active *phoA* fusions, or X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside) for *lacZ* fusions. Incubate in darkness for 2 to 3 days until resistant colonies are developed and average 2 to 3 mm in diameter.

3.2. Arraying

Individual strains were arrayed into 384-well freezer plates using a Qpix colony picking robot (Genetix Ltd.). Using the Qpix, colonies on the bioassay plates (Nunc) were photographed by a grayscale camera, which identifies the colonies and shows the analysis to the user through a graphical interface. These colonies were then picked using a 96-pin picking head. Either blue or white colonies may be picked exclusively during a run by using the grayscale threshold image analysis software that runs the Qpix (**Note 1**). For each run, a negative control was performed to determine if cross-contamination occurred by incomplete pin sterilization during picking. Colonies were arrayed into 384-well plates, each with 80 μL UWGC freezing medium (however, *see* **Note 2**) with 20 μg/mL tetracycline and 10 μg/mL chloramphenicol as follows:

1. Aliquot freezer medium into 80-μL amounts in 384-well freezer plates or PCR-tube strips. Store at 4°C wrapped in Saran Wrap and aluminum foil.
2. Inoculate by toothpick, sterile pipette tip, or Qpix pin.
3. Let grow overnight at 37°C and then place in the −80° freezer for long-term storage (**Note 3**).

3.3. PCR and Sequencing

Protocols for PCR and sequencing were summarized in Jacobs et al. (2). The supplementary methods from that publication contains detailed descriptions of the PCR and sequencing methods, primer sequences, and thermocycler settings and may be accessed at: http://www.genome.washington.edu/UWGC/pseudomonas/pdf/Supplementary_Methods.pdf.

3.4. Mapping: PERL Scripts

Manual determination of the junction point may be accomplished easily by finding the last base of the transposon sequence, trimming off the transposon sequence, and using BLAST to compare the remainder to the *P. aeruginosa* genome. The first 1000 sequencing chromatograms (chromats) from this project were mapped individually and then compared with the automated script described below. Though simple, manual mapping of junctions is time-consuming and is not readily scalable into the 10,000 to 50,000 range.

Automating the analysis of sequence data was readily accomplished by implementation of a custom-designed PERL script. Our model script was written by David Spencer (2), and a basic flow chart is shown in **Figure 3**. This single script can assess the quality of the sequence trace, find the vector-genome junction, map the junction to a reference genome, and reference the map point to annotation data (such as ORF start and stop coordinates and ORF function). Several inputs are necessary: a reference genome, the reference transposon sequence (at least at the junction), and an annotation table for the genome.

A series of simple algorithms are applied to the chromatograms in order to determine the genomic coordinate of the transposon/genome junction at the 5′ end of the transposon. For each chromat, the script uses Phred to determine quality (5, 6) and then exports a text file of the sequence to a defined directory. Cross-match, using the Smith-Waterman algorithm (7), is used to find the last base in the text sequence file to match the transposon sequence. If the last matching base of the chromat corresponds with the last base of the reference transposon sequence, then the junction is called “Exact” (this occurred in 83% of trials). Otherwise, an “Adjusted” call is made (7% of trials), and the script outputs the number of bases missing in the transposon sequence. If the transposon sequence is not found, the output for junction says “None” (10% of trials). For all cases, the script outputs a text sequence file, and when it finds the transposon sequence, it also outputs a “screened” file with the transposon sequence removed. Next, the script uses the screened file, or if none is available, the sequence file, to cross-match against the reference PAO1 genome. If the first base of the sequence to match the genome is also the first base after the transposon sequence, then the genome position call is “Exact” (58% of trials), and a determination about the frame of the inserted fusions may be made. Only insertions with a frame of “+2” are strictly competent to make a translational fusion (sloppy expression of tags is possible and likely happens). If an Exact genome position is not available, then an “Estimate (X)” position is returned, with the parenthetical “(X)”value equal to the number of bases in the chromat between the last base of the transposon match and the first base of the genome

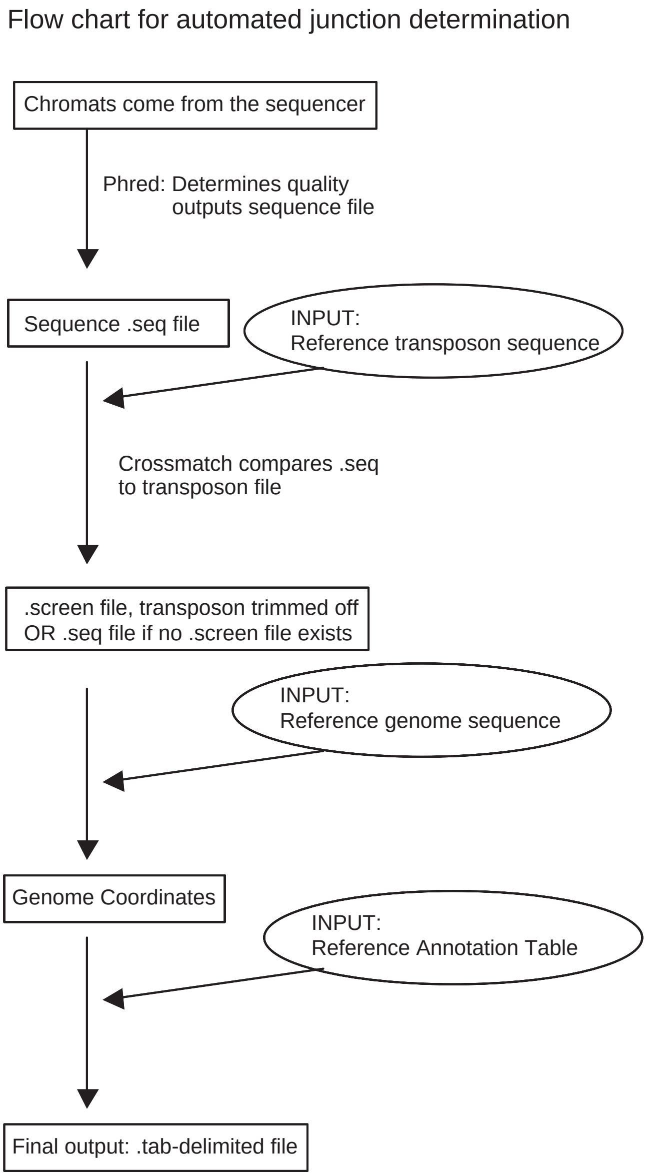

Flow chart for automated junction determination
Chromats come from the sequencer
Phred: Determines quality
outputs sequence file
Sequence .seq file
INPUT:
Reference transposon sequence
Crossmatch compares .seq
to transposon file
.screen file, transposon trimmed off
OR .seq file if no .screen file exists
INPUT:
Reference genome sequence
Genome Coordinates
INPUT:
Reference Annotation Table
Final output: .tab-delimited file

Fig. 3. Automated transposon insertion mapping. A flowchart describing the PERL script used to map transposon insertion, originally written by David Spencer. Parts of this figure were originally published as Figure 4 in Ref. *4* (p. 127) and are reproduced here with kind permission of Springer Science+Business Media.

match. Once a genome coordinate for the insertion has been determined, it is used to retrieve data from the annotation, including: PA ORF number, gene name, gene function (both the long form of the gene name and also the broad functional category as assigned in the annotation table), position relative to the gene in bases and codon number, and the frame of the insertion (as mentioned above). Intergenic hits create a duplicate record with the position relative to the two adjacent ORFs noted in subsequent records.

The data output is in a tab-delimited .txt file, which allows automated import into the database. Log files are generated that tabulate the number of times the transposon was found and the number of successful matches to the genome within a run. These log files are useful for real-time determination of success rate, which is critical during troubleshooting.

3.5. Basic Description of the Produced Mutant Collection

The UWGC *Pseudomonas aeruginosa* library is described in Jacobs et al. *(2)* and in Jacobs and Manoil *(4)*. Physically, one copy of the collection is housed in 110,384-well freezer plates, which are easily housed in two standard 66-plate racks (approximately one-half of a shelf in an upright −80°C degree freezer), corresponding with 42,240 wells (**Note 4**). Several plates were sequenced more than once, and the overall success rate was 80%, success being defined as a genome location identified for the site of transposon insertion. Once failed attempts, siblings, and discrepant positions had been screened out, 30,100 unique insertion locations had been identified.

3.5.1. Candidate Essential Genes

Six hundred seventy-eight ORFs were never hit by a transposon insertion. We designated these ORFs as "Candidate Essentials." As described in Ref. *2* and in **Chapter 22**, statistical analysis suggests that approximately half of these ORFs are likely to be truly essential. Our bioinformatic analysis *(2)* consisted of a BLAST comparison of all *P. aeruginosa* ORFs with those from *E. coli* in the PEC database (http://www.shigen.nig.ac.jp/ecoli/pec/index.jsp; **Chapter 29**). The overall conclusions from the results were that about one-half of the candidate essential genes had a strong homologue to a known essential gene in *E. coli*. This result was consistent with the statistical analysis predicting the total number of essential genes to fall between 300 and 400. Sophisticated bioinformatic analysis is ongoing in collaboration with other laboratories that have developed resources described in this book.

3.6. Database Construction

The most effective method for storing and analyzing large data sets was determined to be a relational database. We selected the Microsoft Access database for ease of setup and the ability to create Visual Basic modules for accelerating analysis and data import. PERL script output is in a tab-delimited text file (see above) and is imported and appended to the main data table via a data capture form that integrates the import and appending functions. Though creating this form is a minor project, its utility is twofold: convenience of data import and elimination if errors caused by manual import. With large data sets, it is less likely that small errors will be noticed.

The most common analysis desired was to determine the number of unique insertion locations. A Visual Basic module was written by Stephen Ernst (UW Genome Center) to accomplish this goal. The module eliminates null records (no sequence match to genome—usually due to sequencing failure), eliminates duplicate records (due to intergenic hits, above), counts the number of "siblings" (identical insertion locations—these may be true siblings or may be due to cross-contamination), and accounts for resequencing duplications. This analysis requires that only one representative from each insertion location be counted, so multiple insertions at the same location are sequestered into a separate "Sibling" table, and only the first representative is kept in the "Unique Inserts" table. Occasionally, resequencing returned discrepant results. In these cases, both records were sequestered to a "Discrepancy" table and were not counted as unique inserts.

3.7. Quality Determination

The main quality determination to be made in a large mutant collection is whether it is verifiable that transposon insertions are accurately mapped to the correct well within a plate of strains. To determine this, we designed custom primers oriented toward the transposon insertion for a randomly selected subset of clones. The general scheme of primer design is given in **Figure 4**. A positive PCR product indicated the presence of a transposon insertion at the mapped position; however, positive results may not identify wells that contain multiple strains and also will not identify strains with multiple transposon insertions. A second primer completing the flanking of transposon insertion point will be informative of whether there are intact genomic segments, which usually is indicative of the presence of another strain in the well.

A total of 112 strains were screened specifically for quality control (QC) purposes, 107 (96%) of which were confirmed by the above analysis. A few strains were also

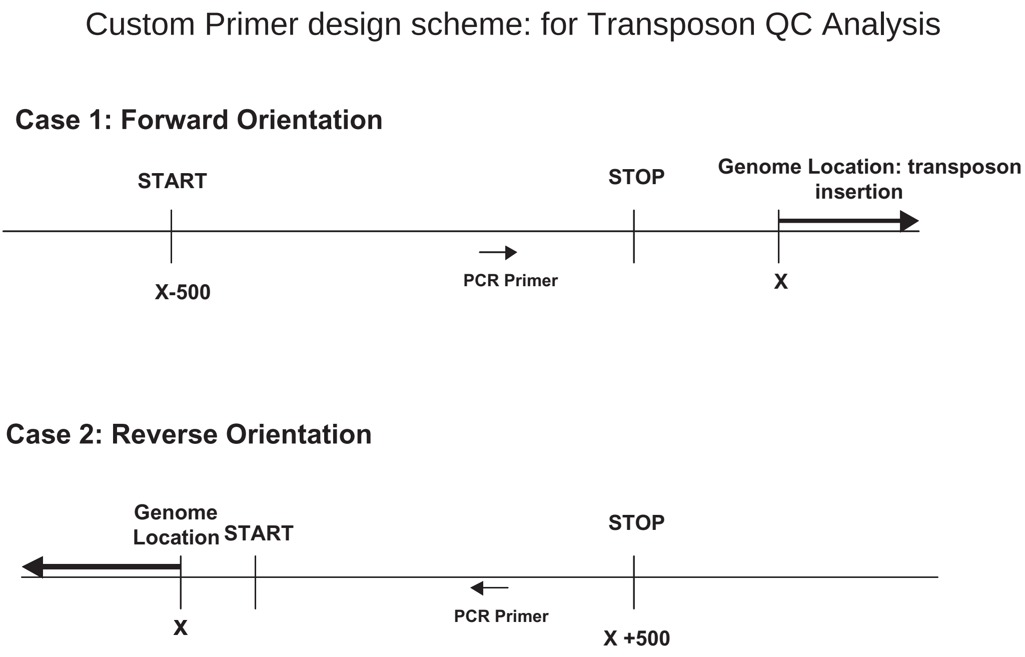

Custom Primer design scheme: for Transposon QC Analysis
Case 1: Forward Orientation
START
STOP
Genome Location: transposon insertion
X-500
PCR Primer
X
Case 2: Reverse Orientation
Genome Location
START
STOP
X
PCR Primer
X +500

Fig. 4. Primer design for PCR confirmation of transposon insertion location. A design scheme is presented for PCR experiments that will confirm the presence of a mapped transposon insertion.

confirmed by flanking PCR, showing that the intact gene has been deleted. Other techniques were used to confirm the identity, including restreaking fusion protein-producing clones on indicator medium. Several recipients of strains have confirmed the mapped position of strains via these methods and also have reported mixed strains in some cases. In many cases, a second mutant in the gene of interest is available for strains where ambiguous QC results occur.

3.8. Strain Maintenance

Strains are maintained in freezer plates stored at −80°C. It is recognized that this is not an absolutely permanent storage solution. We have found that glycerol stocks of *Pseudomonas aeruginosa* lose viability if stored under these conditions over time. Plates may be thawed and replicated using 384-pin replicators. It is prudent to maintain several copies of the library. We maintain three copies at the UW genome center: the original copy, a backup copy (in a separate freezer), and a copy that is used for strain distribution (see below).

Further optimizations of long-term storage protocols are under way, including storage in DMSO stocks, using deep-well freezer plates, and airtight seals on the plates.

3.9. Visual Phenotype Scoring

Any phenotype for which a simple screen is available may be scaled up to 384-well format. Phenotypic analysis was initiated with the primary goal of determining whether the mutant collection would accurately reflect the genomic bases of two known biological processes: twitching motility and growth on minimal medium. We chose two simple visual phenotypes: the lack of twitching motility and no growth on minimal medium. The genes responsible for these two processes have been well characterized in previous studies *(13)*. We recovered insertions in nearly all the genes known to be responsible for these genotypes and concluded that screening the collection for easily visualized mutants yielded results consistent with previous forward-genetic screens. For these phenotypes, a qualitative +/– score is sufficient, although there are occasional subtleties. To score colonies, we used a metal 384-pin replicator to transfer a small subsample from each well of a freezer plate onto solid medium (such as LB agar), poured in plates the same shape as freezer plates. These plates were grown under standardized conditions and then digitally photographed under standardized conditions. Digital photographs were scored manually, and scores were entered into the database according to well number.

Quantitative scoring of growth phenotypes using image analysis tools as have been applied in yeast *(8)* and chemical scoring of clones should be easily adaptable to 384-well format such as that described in Bochner et al. *(9)*. A detailed description of phenotype screening and the lessons learned from this are summarized in Jacobs and Manoil *(4)*. One of the main lessons we learned was that the redundancy of the collection, an average of 5 hits per gene that was hit at least once, was very critical in determination of phenotypic response to mutagenesis for a given gene. In addition, screening for positive phenotypes was far more likely to allow clear determination of genetic correlation.

3.10. Strain Distribution

A primary purpose of the library is to distribute the strains to researchers who may use them individually to assess the biological consequences of mutant phenotypes. Whereas high-throughput phenotype assessment requires a simple screen, the reverse-genetic approach in which candidate genes are screened in more complicated assays, such as animal pathogenesis models, require screening individual clones separately.

Shipping *Pseudomonas aeruginosa*, a class 2 pathogen, is subject to U.S. and international law. A major concern is to keep abreast of changing regulations, including those requiring costly shipping containers. Our standard procedure was to maintain two sets of strains specifically designated for shipping purposes, a "parent" and "working" stock. The parent stock was replicated into the working stock. Each freeze/thaw cycle of the working stock plate was tracked, and after 10 cycles, the parent plate was replicated, and the new replicate was designated as the new parent stock, and the old parent stock became the new working stock. This type of rotating system allowed the smallest number (although still significant) of plate replications. Plate replication in 384-well format is a major concern, due to the close proximity of wells, which makes well-to-well contamination a significant threat. We have been able to retrieve the clone of interest from wells that were contaminated in most cases, with the important exception of wells whose original sequence quality was poor (**Section 3.11**). In any event, the far roomier 96-well format would allow both a more robust volume of glycerol stock to be maintained as well as a greater distance between wells.

We were able to ship nearly 9000 individual strains from the collection along with the parent strain, totaling 9236 stains shipped between November 2003 and February 2006. At that time, our funding to support the strain distribution effort ended. We are currently searching for more support for this effort or for a centralized strain distribution center (such as ATCC) that can handle the shipping.

3.11. Strain Curation

At the time of this writing, we are in the process of curating the collection to the top two strains per ORF. Several criteria were used, including the location of the insertion relative to the ORF (to ensure the likeliest possibility of functional disruption), the orientation and frame of the insertion (to maximize the presence of fusion strains), and the use of the best original data from the high-throughput sequencing. We found that high-quality hits were the most repeatable and the most likely to be retrieved. It is likely that low-quality mapping scores in the original data set could have been due to cross-contamination, multiple insertions per cell, or other random processes. By scoring the insertions for the highest quality of sequence, we have been able to achieve approximately 90% resequencing success rate (unpublished data, M. Jacobs). The curation algorithm was developed by looking individually through the first 1000 genes and picking the top two by hand. Then, a numerical algorithm was written to reproduce those results and was applied to those and the remainder of the ORFs. Development of curation algorithms would likely vary from library to library. Finally, it is also important to keep in mind that the high redundancy of the collection has proved invaluable in recovering phenotypically relevant mutants.

3.12. Conclusion

Using the reference genome as starting material, it has been possible to create a resource that is useful both in genomic applications and in accelerating the rate at which mutants are obtained in candidate genes of interest (**Notes 5, 6,** and **7**). By generating a defined mutant collection, phenotypic consequences of different alleles in the same gene may be followed. Our mutant collection was engineered with transposons that allow excision using the Cre/*loxP* recombination system, which will delete the antibiotic selection conferred by the insertion and leave behind an epitope tag. Thus, double mutants may be made. By using high-throughput technology, it is relatively simple to produce such a collection. Large collections such as this one function importantly as central resources that are able to tie together bioinformatics and experimental techniques.

Notes

1. Blue/white selection and fusion expression: Colonies were plated on indicator medium and were selected as either white or blue colonies. Blue/white selection was accomplished using the image analysis software that controls the Qpix robotic colony picker. It was noticed that occasionally the robot picked colonies of the wrong color, but it was generally accurate, if imperfect. Of the ORFs that received in-frame *phoA* fusions, only 727 of 1125 produced blue colonies. Given that *phoA* fusions generally require secretion for expression of the blue color, it is not surprising that many fusions do not create a blue color. ORFs with in-frame *lacZ* insertions had a roughly even proportion of blue and white colonies (with 49 ORFs producing both colors in different insertions).
2. We have recently switched to plain LB medium, plus 20 μg/mL tetracycline, and 10 μg/mL chloramphenicol, plus *5% DMSO* as a replacement for this freezing medium for *Pseudomonas*. We have found that this new medium works equally well for survival (cultures did not show deterioration after 10 freeze/thaw cycles, and 1 year frozen), and also that the DMSO appears to reduce the "biofilm" nature of many mutant strains (wherein the whole culture in a well will stick to a plastic pipette tip during manipulation). One note: Using 5% DMSO was effective in reducing the biofilm-like phenotype for the collection, which in turn helped reduce the cross-contamination due to large globs of cells clinging to the pin replicators. For *Pseudomonas*, the author would highly recommend DMSO over glycerol.
3. For high-throughput analysis, it is critical to carefully track plates to make sure that the correct plate is tied together with sequencing results downstream. As opposed to a shotgun-sequencing project, each well must be correctly associated with all sequencing reads that originate from it, or its downstream value for recovering individual clones or phenotyping will be lost.
4. The decision of when to stop mapping inserts is based on determining how many hits in new ORFs will be returned when searching 384 new strains (**Chapter 22**). Sequencing the 110th plate of mutants, we saw only about 30 or fewer new ORFs hit per 384 trials. In addition, we had an average of 5.75 hits per ORF that had been hit at least once, and 5.05 hits on average per every predicted ORF in the genome.
5. Use of the collection to define gene function: One of the more interesting results of producing this mutant library has been watching which categories of mutants are most popular. When the strains sent are categorized by functional category in **Table 3**, we see that a lot of

Table 3
Changes in Annotation in Requested Strains

Putative ORF function	Functional class 2000	Functional class 2006
Adaptation, protection	260	630
Amino acid biosynthesis and metabolism	224	297
Antibiotic resistance and susceptibility	115	36
Biosynthesis of cofactors, prosthetic groups	45	207
Carbon compound catabolism	118	85
Cell division	6	9
Cell wall/LPS/capsule	256	345
Central intermediary metabolism	79	101
Chaperones and heat shock proteins	77	13
Chemotaxis	309	102
DNA replication, recombination, modification	275	245
Energy metabolism	211	236
Fatty acid and phospholipid metabolism	114	64
Hypothetical, unclassified, unknown	2303	1560
Membrane proteins	57	1463
Motility and attachment	430	361
Nucleotide biosynthesis and metabolism	30	56
Protein secretion/export apparatus	210	201
Putative enzymes	472	455
Related to phage, transposon, or plasmid	112	87
Secreted factors	392	419
Transcription, RNA processing, and degradation	10	11
Transcriptional regulators	912	1015
Translation, posttranslational modification	143	110
Transport of small molecules	1136	522
Two-component regulatory systems	677	343

Many of the strains that have been requested have had changes to the annotation in their associated ORFs in the time period between 2000 and 2006. For example, 2303 strains have been sent (including duplicated strains to different researchers and multiple mutants within the same gene) that were annotated as unclassified in 2000, but only 1560 of those strains would still have their parent ORFs unannotated. Of the 8973 mutant strains sent since November 2003, 3550 strains are in genes that have had updates to their annotations at least equivalent to changing the functional class for the ORF.

the action in terms of updating the annotation table corresponds closely with the strains we have shipped. Many of the strains shipped correspond with mutants in genes that have had their functions further defined in the time between the original annotation and the current annotation as of May 2006. Much of this change has been in the past 2 years. Different functional classes of genes remain highly differential in their popularity among researchers, as shown in **Table 4**. Exported genes are particularly differential, especially as they relate to host-recognition and the resulting immune response functions *(8)*. Of the individual ORFs,

Table 4
Popular Functional Class Categories

Function class	Strains shipped	Strains available	Relative proportion
Chemotaxis	102	120	85%
Secreted factors	419	502	83%
Motility and attachment	361	445	81%
Adaptation, protection	630	838	75%
DNA replication, recombination	245	410	60%
Cell wall/LPS/capsule	345	584	59%
Protein secretion/export apparatus	201	346	58%
Two-component regulatory systems	343	649	53%
Antibiotic resistance and susceptibility	36	71	51%
Transcriptional regulators	1015	2089	49%
Biosynthesis of cofactors, prosthetic	207	555	37%
Membrane proteins	1463	5080	29%
Related to phage, transposon, or plasmid	87	339	26%
Transport of small molecules	522	2201	24%
Amino acid biosynthesis and metabolism	297	1372	22%
Nucleotide biosynthesis and metabolism	56	259	22%
Energy metabolism	236	1129	21%
Fatty acid and phospholipid metabolism	64	309	21%
Translation, posttranslational modification	110	583	19%
Chaperones and heat shock proteins	13	75	17%
Cell division	9	52	17%
Central intermediary metabolism	101	606	17%
Putative enzymes	455	2858	16%
Hypothetical, unclassified, unknown	1560	10,491	15%
Carbon compound catabolism	85	611	14%
Transcription, RNA processing	11	263	4%

Of strains available for shipping within a category, a relative difference between the numbers available shipped exists per category. Of the 120 strains available for genes involved in chemotaxis, 102 were shipped at least once. At the other extreme, central cell metabolism genes were not requested. Presumably, it is the cell-surface/antigenic properties of exported genes that make them more popular.

the most-requested ones often do not fall into the most-requested categories, and one can surmise that this is due to intense interest in these genes' individual functions (**Table 5**).

6. Synergy and comparison in multiple collections: Three large collections have now been produced for *Pseudomonas aeruginosa* (***2, 9, 10***). Between the three, there are only 458 ORFs that were not hit in any collection (**Table 6**). The power of integrating collections is further illustrated in the GBrowse genome annotation tools developed and available at www.pseudomonas.com. The utility of different collections in different clinical strains will likely greatly leverage the power of any individual collection.
7. Publications citing our collection: Finally, the utility of our collection may be judged by the number of publications where the collection was cited. Since July 2004, the trend has been upward (**Fig. 5**).

Table 5
The Most Popular *P. aeruginosa* Genes in the World

PA ORF		Name Product	Strains shipped
PA1003	mvfR	Transcriptional regulator	55
PA3477	rhlR	Transcriptional regulator RhlR	51
PA4525	pilA	Type 4 fimbrial precursor PilA	51
PA3724	lasB	Elastase LasB	48
PA0996	pqsA	Probable coenzyme A ligase	46
PA1092	fliC	Flagellin type B	40
PA0652	vfr	Transcriptional regulator Vfr	37
PA3622	rpoS	Sigma factor RpoS	37
PA5368	pstC	Membrane protein component of ABC phosphate transporter	35
PA4700	mrcB	Penicillin-binding protein 1B	34
PA5367	pstA	Membrane protein component of ABC phosphate transporter	32
PA4110	ampC	Beta-lactamase precursor	27
PA1871	lasA	LasA protease precursor	27
PA3841	exoS	Exoenzyme S	27
PA4633		probable chemotaxis transducer	26
PA3478	rhlB	Rhamnosyltransferase chain B	25
PA3790	oprC	Putative copper transport outer membrane porin OprC precursor	25
PA3617	recA	RecA protein	25
PA1001	phnA	Anthranilate synthase component I	25
PA1180	phoQ	Two-component sensor PhoQ	25
PA1777	oprF	Major porin and structural outer membrane porin OprF precursor	25

The number of strains shipped from the UWGC collection organized by parent gene, sorted by the most strains sent.

Table 6
Candidate Essential Genes

Primary Function (Annotation 2006)	UW not-hit	Pathogenesis hits	PA14 hits	Not hit in any collection
Adaptation, protection	5	1	1	4
Amino acid biosynthesis and metabolism	30	11	8	21
Antibiotic resistance and susceptibility	1	1	1	0
Biosynthesis of cofactors …	41	10	12	31
Carbon compound catabolism	16	11	13	5
Cell division	10	1	4	9
Cell wall/LPS/capsule	25	1	2	24
Central intermediary metabolism	11	3	4	8
Chaperones and heat shock proteins	4	1	1	3
DNA replication, recombination	16	0	3	16
Energy metabolism	31	10	13	21
Fatty acid and phospholipid metabolism	14	3	3	11
Hypothetical, unclassified, unknown	235	86	119	149
Membrane proteins	42	19	27	23
Motility and attachment	1	0	1	1
Nucleotide biosynthesis and metabolism	13	2	3	11
Protein secretion/export apparatus	19	8	5	11
Putative enzymes	31	17	14	14
Related to phage, transposon, or plasmid	9	4	1	5
Secreted factors	7	3	5	4
Transcription, RNA processing	12	1	6	11
Transcriptional regulators	32	20	18	12
Translation, posttranslational	62	9	15	53
Transport of small molecules	12	2	0	10
Two-component regulatory systems	2	1	2	1
Totals	**681**	**225**	**281**	**458**

Candidate essentials are defined as ORFs for which a mutant was not obtained in a saturation or near-saturation mutant collection. In the UWGC collection, 681 ORFs did not have a hit internal to the open reading frame. However, out of those, 225 were hit in the Pathogenesis collection (unpublished data), and 281 homologues to those were hit in the PA14 collection (**Chapter 7**; http://ausubellab.mgh.harvard.edu/cgi-bin/pa14/home.cgi). 458 ORFs were not hit in any of the collections (**Note 6**). In the UWGC collection, there were additional 106 genes that were only hit in the last 10% of their open reading frames. Overall, 4892 PAO1 ORFs were hit in the UWGC collection, 3908 were hit in the pathogenesis collection, and 3725 PAO1 homologues are included in the PA14 collection.

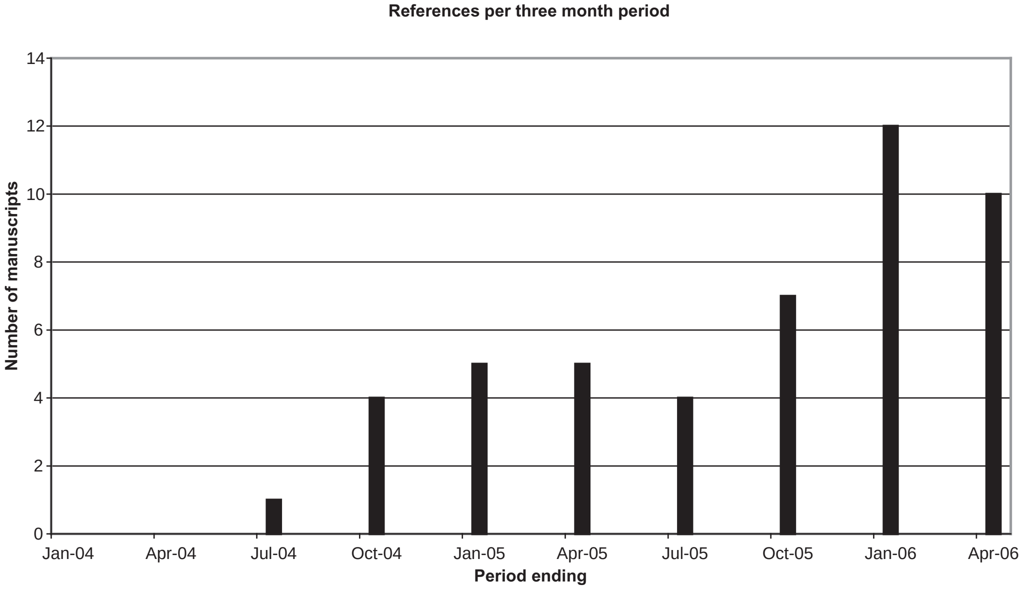


Fig. 5. Citations for the UWGC mutant library. The number of citations grouped into 3-month time periods since July 2004 has been increasing. Citations are still occurring (**Note 7**).

Acknowledgments

The author would like to thank Dr. Maynard Olson and Dr. Colin Manoil for initiating the project; Elizabeth Sims for managing the worldwide strain distribution; Dr. Sam Miller for his continued support; Amy Dao, Stephen Ernst, and Gregory Alexander for organizational and informatics support for strain distribution; and Drs. Shawn Lewanza and Robert Hancock for helping to redefine our freezing medium. This work was supported by grants from the Cystic Fibrosis Foundation.

References

1. Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., et al. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406**, 959–964.
2. Jacobs, M. A., Alwood, A., Thaipisuttikul, I., Spencer, D., Haugen, E., Ernst, S., et al. (2003) Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 14339–14344.
3. Bailey, J., and Manoil, C. (2002) Genome-wide internal tagging of bacterial exported proteins. *Nat. Biotech.* **20**, 839–842.
4. Jacobs, M. A., and Manoil, C. (2006) A genome-wide mutant library of *Pseudomonas aeruginosa*. In: Ramos, J. L. and Levesque, R., eds. *Pseudomonas, Volume 4*. Dordrecht, The Netherlands: Springer, pp. 121–138.
5. Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998) Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185.
6. Ewing, B., and Green, P. (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* **8**, 186–194.

7. Smith, T. F., and Waterman, M. S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
8. Fogel, G. B., and Brunk, C. F. (1998) Temperature gradient chamber for relative growth rate analysis of yeast. *Anal. Biochem.* **260**, 80–84.
9. Bochner, B. R., Gadzinski, P., and Panomitros, E. (2001) Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Methods* **11**, 1246–1255.
10. Wu, L., Estrada, O., Zaborina, O., Bains, M., Shen, L., Kohler, J. E., et al. (2005) Recognition of host immune activation by *Pseudomonas aeruginosa*. *Science* **309**, 774–777.
11. Liberati, N. T., Urbach, J. M., Miyata, S., Lee, D. G., Drenkard, E., Wu, G., et al. (2006) An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2833–2838.
12. Bruce, K. (2005) Personal communication.
13. Semmler, A. B., Whitchurch, C. B., and Mattick, J. S. (1999) A re-examination of twitching motility in *Pseudomonas aeruginosa*. *Microbiology* **145**, 2863–2873.

10

Comparing Insertion Libraries in Two *Pseudomonas aeruginosa* Strains to Assess Gene Essentiality

Nicole T. Liberati, Jonathan M. Urbach, Tara K. Thurber, Gang Wu, and Frederick M. Ausubel

Summary

Putative essential genes can be identified by comparing orthologs not disrupted in multiple near-saturated transposon insertion mutation libraries in related strains of the same bacterial species. Methods for identifying all orthologs between two bacterial strains and putative essential orthologs are described. In addition, protocols detailing near-saturation transposon insertion mutagenesis of bacteria are presented, including (1) conjugation-mediated mutagenesis, (2) automated colony picking and liquid handling of mutant cultures, and (3) arbitrary polymerase chain reaction amplification and sequencing of genomic DNA adjacent to transposon insertion sites.

Key Words: essential genes; *MAR2xT7*; *mariner*; PA14, *Pseudomonas aeruginosa*.

1. Introduction

The availability of multiple, nearly saturated mutant libraries in related strains of a single bacterial species offers the opportunity of identifying orthologous genes that are nondisrupted in more than one library. The set of nondisrupted genes are putative "essential" genes. In this chapter, methods are described for creating a nearly saturated bacterial transposon insertion library including conjugation-mediated mutagenesis, arraying transposants into plates using a colony-picking robot, and aliquoting mutant cultures using a liquid-handling robot. Specific guidelines, based on quality control testing, are described for automated handling of bacterial cultures that minimize cross-contamination. Methods are also described for identifying transposon insertion sites using two-step polymerase chain reaction (PCR) amplification of the DNA adjacent to the transposon insertion site using arbitrary primers and subsequent sequencing of the PCR products.

This chapter also describes the use of a custom-designed database for automated DNA sequence analysis. DNA sequences adjacent to transposon insertions are entered into the database where they are aligned with the genomic sequence of the mutagenized

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

strain using the BLAST algorithm. The approximate genomic locus of each insertion site is determined by the alignment with the best BLAST score. The precise location of each insertion is determined using a modified Smith-Waterman algorithm that aligns sequences obtained from each mutant with the 3′ end of the transposon sequence. Once all insertion sites have been located, genes that have been disrupted as well as those that have not can be identified. Finally, protocols are described that identify orthologs in two bacterial strains that can be used to detect essential genes based on the absence of insertion mutants in the orthologs in more than one transposon mutant library.

The protocols described in this chapter are based on experiments carried out in our laboratory that were involved in the use of the *mariner*-based transposon *MAR2xT7* to generate a mutation library in *Pseudomonas aeruginosa* strain PA14 (*1*). Nevertheless, many if not all of the protocols can be readily adapted for generating transposon mutation libraries in most Gram-negative bacterial species.

2. Materials

2.1. Bacterial Strains

1. Recipient strain: the strain to be mutagenized must be λ *pir*-.
2. Donor strain: *pir*$^{+}$ *Escherichia coli* strain carrying plasmid containing a transposase, a compatible transposon that confers antibiotic resistance, additional antibiotic resistance marker outside of the transposon, and a bacteriophage λ *pir*-dependent origin of replication. In the case of the *P. aeruginosa* strain PA14 library (*1*), the vast majority of mutants were created with *MAR2xT7*, a gentamicin-resistant derivative of the *Himar1* transposon (*2, 3*). All protocols in this chapter are based on *MAR2xT7* insertion.
3. Helper strain: If the donor plasmid carrying the transposon is mobilizable but not self-transmissible (*mob*+ *tra*-), an *E. coli* strain carrying a broad-host range helper plasmid should be included in the mating to facilitate conjugation. For example, pRK2-derived IncP broad-host range plasmids encode the *tra* genes necessary for conjugal transfer in *trans* (*4*). If using a helper plasmid, the donor plasmid carrying the transposon requires a proper mobilization sequence known as the origin of conjugational replication (*5*). To mobilize the donor plasmid encoding *MAR2xT7*, we used a helper strain, *E. coli* HB101 carrying the pRK2 derivative pRK2013, in all matings.

2.2. Relational Database

1. A relational database in needed to (a) track mutant location, (b) track processing status information, and (c) analyze sequencing data for each mutant. The database must contain the genomic sequence of the strain being mutagenized and the coordinates of all predicted open reading frames (ORFs). The PA14 Transposon Insertion Mutant Database (PATIMDB) that was developed and used in our laboratory is implemented using the MySQL RDBMS hosted on a multiprocessor Intel system running RedHat Linux (*1*). The data-entry application is in Java and runs on Windows 2000. PATIMDB is compatible with different genome sequences and is adaptable to library construction applications in other organisms. A "generic" version of PATIMDB that is designed for use with any bacterial genome will be downloadable in the future at http://ausubellab.mgh.harvard.edu/cgi-bin/pa14/downloads.cgi.

2.3. Equipment

1. Q-fill media dispenser (Genetix, Boston, MA).
2. QBot colony-picking robot (Genetix).
3. Biomek FX liquid-handling robot (Beckman Coulter, Inc., Fullerton, CA).
4. HiGro block shaker/incubator (Genomic Solutions, Ann Arbor, MI).
5. Thermocylers with capacity to run multiple 96-well plates in parallel (e.g., ThermoHybaid, Ashford, Middlesex, UK).
6. ABI 3700 PRISM automated sequencer.
7. 12-channel pipettes (Costar or Finnpipette) with compatible tips.
8. Tabletop centrifuges with block/plate attachments (Beckman Coulter Allegra X-22, Fullerton, CA).
9. Laminar flow hood.
10. –80°C and –20°C freezers including racks designed to hold 96-well plates.
11. Aluminum bases to cool 96-well plates.

2.4. Reagents

1. Arbitrary PCR and sequencing primers (*see* **Methods** and **Fig. 1**).
2. Deoxynucleotide triphosphates (dNTPs), PCR grade (Roche, Indianapolis, IN).
3. Taq DNA polymerase and 10× buffer (no. 1147633; Roche).

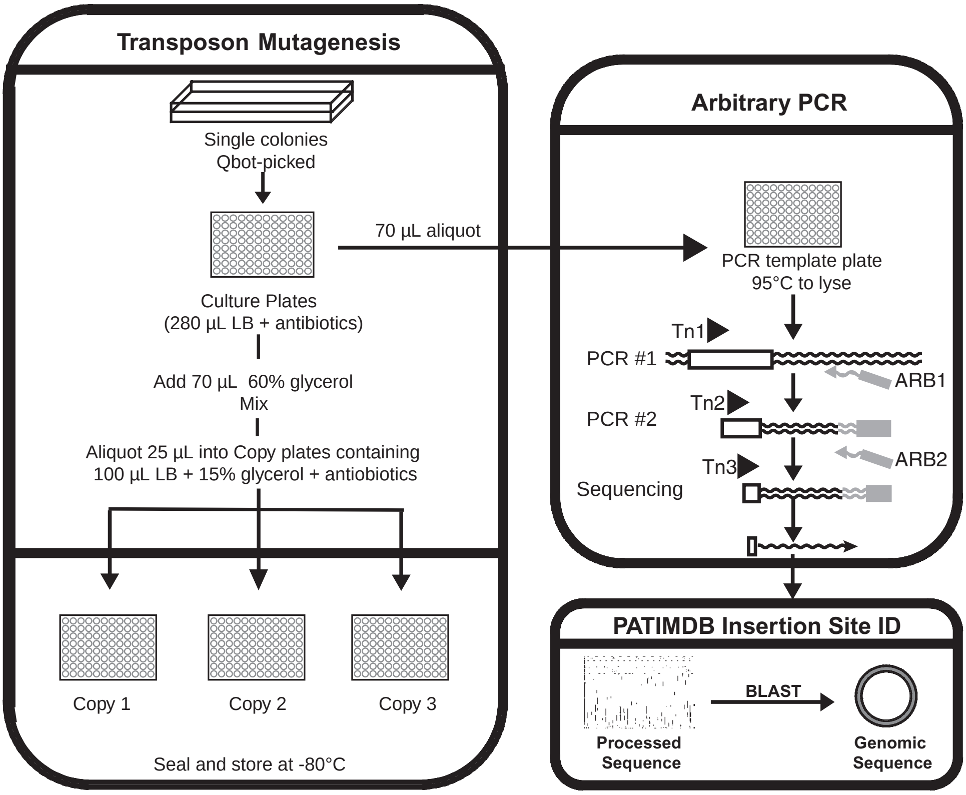


Fig. 1. Transposon insertion library production and analysis workflow.

4. Dimethyl sulfoxide (DMSO) (no. D-8418; Sigma-Aldrich, St. Louis, MO).
5. ExoSAPIT (no. 78205; USB Cleveland, OH).
6. Applied Biosystems Taq Dye Deoxy Terminator cycle sequencing kits.
7. King's B media (2% w/v peptone, 6.57 mM K_2HP0_4, 6.08 mM $MgSO_4$, 1% v/v glycerol).
8. Luria Broth (1% w/v peptone, 0.5% w/v yeast extract, 0.5% w/v NaCl).
9. Sterile 60% glycerol.
10. Antibiotics to select for the transposon and to select against other mating strains.

2.5. Consumables

1. 20 cm × 20 cm low-profile bioassay dish (no. 240845; Nunc, Rochester, NY).
2. Glass balls 3 mm in diameter (no. 26396-508; VWR, West Chester, PA).
3. Biomek AP96 P250 tips or equivalent (no. 717251; Beckman Coulter, Fullerton, CA).
4. Culture plates/blocks: Greiner 96-well flat-bottom plates with lids (Greiner no. 655185; or alternatively Culture Blocks: 2.0 mL 96-well V-bottom polypropylene blocks (no. 3961; Costar/Corning, Acton, MA).
5. Sealing mats for 2.0-mL 96-well blocks (no. 3083; Costar/Corning).
6. Copy plates: low-profile 96-well Serowel V-bottom plates (Bibby Sterilin Serowel no. 611V96 or Abgene no. MP-2000).
7. Lids with stacking rim for 96-well Serotec plates (no. AB-0752; Abgene, Rochester, NY).
8. PCR template plates: 96-well skirted thermo-fast reaction plates (no. AB-0800/150; Abgene).
9. PCR reaction plates: 96-well thin-walled, skirted polycarbonate PCR plates (no. 6511; Costar/Corning).
10. Sterile ARB reaction mix reservoirs (no. 13681501; Fisher Scientific, Pittsburgh, PA).
11. Aeroseal breathable seals (no. B-100, Excel Scientific, Wrightwood, CA).
12. AluminaSeal temperature-resistant seals (no. ALUM-1000; Diversified Biotech Boston, MA).
13. Temperature-resistant cryo-tags (plate labels) (no. SIDE-1000; Diversified Biotech).

3. Methods

3.1. Mutagenesis and Arraying Mutants into 96-Well Plates

If the donor plasmid carrying the transposon is self-transmissible, the strain carrying the donor plasmid is directly mated with the recipient strain. If, however, the donor plasmid is not self-transmissible but is mobilizable, triparental mating with donor, recipient, and helper strains should be performed.

3.1.1. Mating (Note 1)

1. Grow separate saturated cultures of transposon donor and recipient strains (and if necessary, helper strain) with appropriate antibiotics in appropriate media (e.g., Luria Broth; LB).
2. Mix 200 µL recipient strain culture with 400 µL donor strain culture (and, if necessary, 400 µL helper strain culture). Gently pellet cells. Generally a 2 : 1 ratio of donor to recipient culture volume is recommended but should be tailored to individual mating combinations. It may be necessary to set up multiple mating mixes depending on the frequency of transposition and mating efficiency.

3. Rinse pellet in 1 mL mating media. For PA14, King's B media was used (*6*). However, other media, including LB, may be appropriate for other mating combinations. Gently pellet cells. Resuspend pellet in 250 µL King's B media.
4. Spot 25-µL aliquots of the resuspended mixture on King's B media 1.5% agar plates, keeping the area of the drops on the plates as small as possible. Let plates dry before moving to 37°C incubator. Incubate plates for shortest time necessary for transposition (**Note 2**).
5. Using a sterile pipette tip, scrape one or more mating spots into a tube containing 48 mL 0.1 M $MgSO_4$. Resuspend thoroughly by vortexing vigorously. The number of spots that you need to resuspend depends on the frequency of transposition and mating efficiency (**Note 3**).
6. Using approximately 30 sterile glass balls, spread 1.5 mL of the suspension on Luria broth agar (1.5% w/v) in separate 20 cm × 20 cm bioassay dishes containing appropriate antibiotics. To select for transposants, include the antibiotic that the transposon confers resistance to. To select against the donor strain, include an antibiotic to which the recipient strain is resistant to but the donor strain is not (**Notes 4** and **5**). Be sure to spread the culture evenly across the plates. For compatibility with the QBot colony picking robot, the 20 cm × 20 cm dishes should contain exactly 200 mL media.
7. Dry dishes for approximately 45 min in a laminar flow hood until all fluid has dried. Incubate dishes at 37°C for 12 to 15 h.
8. Store dishes at 4°C prior to colony picking.

3.1.2. Label Plates

Create "virtual" plates in the database representing culture plates. The database program should automatically create a set of unique identifiers for those plates. The database should also create "virtual" mutants with their own unique identifiers that are linked to culture plate well positions so that sequencing data for individual mutants can be properly entered and stored. Unique identifiers are numbers or alphanumeric keys that unambiguously specify a particular database entity including plates, mutants, and so forth. Finally, the database should also create human-readable numeric plate labels and, if desired, bar-code labels encoding unique identifiers for each culture plate. These labels should be generated as text files that can be printed (*see* below).

Optional: In the generation of the PA14 transposon library, PATIMDB-created virtual plates for culture, PCR template, PCR reaction, and copy plates, thereby allowing the status of each plate (e.g., whether the ARB 1 reaction has been performed) to be entered and monitored. This tracking feature is not essential but may be useful with different applications. Files containing bar codes encoding unique identifiers and human-readable text labels for each culture plate and the PCR template plates, PCR reaction plates, and copy plates derived from each culture plate are automatically generated by the database once virtual culture plates have been created.

1. Print and apply labels. Files containing bar codes and/or corresponding human-readable labels can be created with Sagian Print and Apply Software (part no. 148640; Sagian Core Systems, Indianapolis, IN). The labels are then printed and transferred to each plate via a print-labeling machine. If this software/hardware is not available, labels for each plate generated by the database can be printed on special temperature-resistant labels (**Materials**) with general word-processing software and a standard laser printer and applied

to each plate by hand. Labels should be applied to one of the two short sides of the plates because these sides face out in the racks used to store plates at −80°C.

3.1.3. Robotic Colony Picking and Inoculation of Culture Plates

The QBot colony picking robot comes with a computer workstation and custom software (QSoft) to run the robot (**Note 6**).

1. Load QBot Platform. Place 20 cm × 20 cm dishes (up to four at a time) containing transposant colonies on the QBot platform. Fill ethanol bath with 70% ethanol.
2. Start QBot software.
3. Camera alignment: Make a hole with a pipette tip in a colony-free area of the agar media in dish no. 1 (the dish located in the upper-left-hand holder on the QBot platform). A dish containing no colonies can also be used. Choose "Align Camera," select "Assay Tray," and then click "Yes." Change "Zoom Focus" to 3 and, using the cursors, set the bull's-eye on the hole. This procedure sets X-Y coordinates for the picking head.
4. Set picking height. Mark "Stop Short on Z Axis," and click the center dot on the screen. Select "Pin A1 Down" over dish no. 1. Click the down arrow, carefully directing the pin to move down until the pin just touches (but does *not* pierce) the agar. Enter "OKAY" to exit. Repeat setting picking height for each of the four dishes on the platform. Return to home position.
5. Test imaging colonies. Each plate is divided into 40 sectors. Select sector 18 of dish no. 1 (the sector located in the middle of dish no. 1) by highlighting it. The camera will move to sector 18 of dish no. 1. Take image by selecting "Picture." A picture of the selected sector will appear with well-defined colonies in green, unclear colonies in yellow, and poor colonies in red. Select "Tools" and "Threshold" to adjust the light settings to maximize the green-to-red colony ratio (**Note 7**). Click "Reprocess." Check other sectors in the same tray to ensure that the green-to-red colony ratio is high. If it is not, return to sector 18 and readjust light threshold as necessary. Check other trays and determine if the set light threshold can be used across all plates on the platform. The light threshold can only be adjusted in sector 18 for each plate. Select "Done" when all trays have been imaged.
6. Set picking run. The software will ask if the steps outlined above have been completed. Select "OKAY" for each of these questions. Specify whether all sectors in all trays are to be picked (full run) or only specific sectors in specific trays (partial run). If a partial run is selected, highlight which sectors should be picked. Once the imaging is complete, a message will appear: "Script is Complete." View image result in the last picked sector screen, which will show the total colonies identified in the run (green, yellow, and red) and the total colonies that the QBot would pick (green only) based on the set parameters in all sectors in every dish.
7. Select "OKAY" to view the Destination Plate Guide. Based on the number of colonies that will be picked as described above, the plate guide will indicate the number of destination plates and where to load them in the two hotels on the QBot platform.
8. Prepare destination plates. Use the QFill media dispenser to fill labeled 96-well flat-bottom plates with 280 μL LB containing appropriate antibiotics to select for transposants. Replace lids and load into the QBot's hotels with the cut corner side of the plate positioned outward. If using different plates or media volumes, be sure the media level in the wells is high enough so that the QBot head pins actually touch the media. Select "Done" to exit.
9. Begin picking. A message will appear: "Are you ready to begin?" Select "YES." The pins are rinsed in 70% ethanol and dried for 10 s between each plate. If more destination plates

are needed, the QBot head will stop in front of the last picked plate in the hotel, and the software will indicate how many more plates to reload. After the picking run, the message “Picking is Complete” will appear. Select “OKAY.” If you want to have the QBot pick from additional trays, select “YES” when the “Do you want to save?” message appears. The QBot will resume picking, inoculating the remaining noninoculated wells, if any, in the last destination plate used (**Note 8**).

10. Carefully seal destination plates with Aeroseal covers, taking care not to disrupt the media.

3.1.4. Culture Plate Incubation

Incubate the inoculated and sealed Culture plates at 37°C without shaking long enough to ensure that slow-growing mutants produce cultures (**Note 9**).

3.1.5. Aliquoting Mutant Strains Using a Biomek FX Liquid-Handling Robot (Fig. 1)

The Biomek FX transfers strains from culture plates to several destination plates: a PCR template plate and three copy plates. This can be done by hand using multichannel pipettes but would require extensive labor. The following steps are used to set a Biomek FX program. Once all plates are properly loaded onto the Biomek platform with lids removed, run the program.

3.1.5.1. FX Method 1: Library Replication for Storage and PCR Processing (Notes 10, 11, and 12)

1. Label and QFill copy plates with 100 μL LB containing 15% glycerol and antibiotics (**Note 13**).
2. Set up deck: Two boxes of tips, one set in the home position; one reservoir filled with sterile 60% glycerol; a culture plate containing fully grown cultures; a bar-coded/labeled PCR template plate; and three bar-coded/labeled and QFilled 96-well copy plates.
3. Pick up tip set no. 1.
4. Aspirate 70 μL of each culture from culture plates at 5 mm below the liquid surface, touch tips to side of wells, and transfer to labeled PCR template plates (**Notes 10, 11,** and **12a**). Touch tips to side of wells (**Note 12b**). PCR template plate cultures are subsequently used as templates for arbitrary PCR.
5. Discard tips.
6. Move tip set no. 2 into the home position.
7. Pick up tip set no. 2.
8. Aspirate 70 μL 60% sterile glycerol from the reservoir 2 mm from the bottom of the reservoir at 70% speed and transfer to the culture plate. Dispense glycerol 5 mm below the liquid surface at 70% speed, with tips following the liquid level as it rises (**Notes 10** and **11**).
9. Mix at least three times 100 μL at 70% speed, following the liquid level as it rises and falls with each aspiration and dispense step with no blowout. Final glycerol concentration after mixing is 15% glycerol. Other types of bacteria may require different glycerol concentrations or other freezing agents such as DMSO. Adjust as necessary.
10. Using the same tips (or a fresh set of tips if desired) aspirate 25 μL culture/15% glycerol mix from 5 mm below the liquid surface, following liquid level, and touch tips to wells. Transfer culture/glycerol mix to the first labeled 96-well copy plate and dispense 3 mm

from the copy plate bottom at 70% speed. Mix 1 μL once at 100% speed (to remove hanging culture drop) with no blowout. Include a tip touch.

11. Repeat **step 10** and dispense in the second copy plate.
12. Repeat **step 10** and dispense in the third copy plate.
13. Seal all plates with AluminaSeals.
14. Store culture and copy plates at −80°C. Store PCR template plates at −20°C until PCR processing.

3.2. Insertion Site Identification

3.2.1. Arbitrary PCR

Transposon insertion sites were identified using a two-round arbitrary PCR protocol (**Fig. 1**) *(9)*.

3.2.1.1. The First Round of Arbitrary PCR (ARB1 Reaction)

1. Thaw PCR template plates containing 70 μL aliquots of the statically grown transformant cultures (*see* above) and incubate at 95°C for 10 min to lyse the cells.
2. Pellet debris by centrifuging the plate at 3000 rpm for 5 min. The cleared lysate is used as template for the first round of arbitrary PCR (ARB1).
3. Combine reagents for the ARB1 reaction mix in a sterile tube on ice: 1× Taq buffer (Roche), 10% DMSO, 2.5 μM dNTPs (**Note 14**), 1.25 U Taq DNA polymerase (Roche), 1.0 ng/μL of the transposon specific primer Tn1 (**Fig. 1**), and an arbitrary primer (**Table 1**) (**Note 15**). Taq is added after all other reagents are mixed thoroughly. Once Taq is added, mix by inverting the tube. For *P. aeruginosa* PA14 *MAR2xT7* mutants, the transposon-specific primer, PMFLGM.GB-3a, (5′-TACAGTTTACGAACCGAACAGGC-3′) was used. Transfer ARB1 reaction mix to a reservoir on ice that will accommodate a 12-channel pipettor.
4. Using a 12-channel pipette, transfer 25 μL ARB1 reaction mix to the wells of a thin-walled 96-well PCR reaction plate sitting on ice in an aluminum plate cooler.
5. Using a 12-channel pipette, transfer 3 μL of the cleared lysates to the reaction mix, pipetting up and down three times to mix.
6. Seal plates with adhesive foil seals.
7. Begin the ARB1 reaction program on the thermocycler:
   (a) 95°C for 5 min
   (b) 95°C for 30 s
   (c) 47°C for 45 s
   (d) 72°C for 1 min
   (e) Repeat **steps** (**b**), (**c**), and (**d**) for 30 cycles.
   (f) 72°C for 5 min
8. Once the thermocycler reaches 95°C for initial denaturation, transfer the ARB1 reaction plate to the thermocycler.
9. Update the status of plates processed for the ARB1 reaction in the database if desired.

3.2.1.2. The ARB2 Reaction

1. Combine reagents for the ARB2 reaction mix in a sterile tube on ice: 1× Taq buffer (Roche), 10% DMSO, 2.5 μM dNTPs, 1.25 U Taq polymerase (Roche), 1.0 ng/μL of the transposon specific primer Tn2 (**Fig. 1**), and an arbitrary primer (either ARB2 or ARB2A; **Table 1**). As in the ARB1 reaction, Taq is added after all other reagents are mixed

Table 1
Arbitrary Primers

First round		Second round	
Primer name	Sequence	Primer name	Sequence
ARB1	GGCCACGCGTCGACTAGTACNNNNNNNNNNGATAT	ARB2	GGCCACGCGTCGACTAGTAC
ARB1A	GGCCACGCGTCGACTAGTACNNNNNNNNNNGTATA		
ARB1B	GGCCACGCGTCGACTAGTACNNNNNNNNNNACNG		
ARB1C	GGCCACGCGTCGACTAGTACNNNNNNNNNNGTAT		
ARB1D	GGCCAGGCCTGCAGATGATGNNNNNNNNNNGTAT	ARB2A	GGCCAGGCCTGCAGATGATG
ARB1E	GGCCAGGCCTGCAGATGATGNNNNNNNNNNGTANG		

thoroughly. Once Taq is added, mix by inverting the tube. For *MAR2xT7* mutants, the transposon-specific primer, PMFLGM.GB-2a, (5′-TGTCAACTGGGTTCGTGCCTT CATCCG-3′) was used. Transfer ARB2 reaction mix to a reservoir on ice that will accommodate a 12-channel pipettor.

2. Using a 12-channel pipette, transfer 20 µL ARB2 reaction mix to a thin-walled PCR reaction plate sitting on ice in an aluminum plate cooler.
3. Using a 12-channel pipette, transfer 5 µL of each ARB1 product to the ARB2 reaction mix in the PCR plate, pipetting up and down three times to mix.
4. Seal plates with adhesive foil seals.
5. Begin the ARB2 reaction program on the thermocycler:
   (a) 95°C for 30 s
   (b) 45°C for 45 s
   (c) 72°C for 1 min
   (d) Repeat **steps** (**a**), (**b**), and (**c**) for 40 cycles
   (e) 72°C for 5 min
6. Once the thermocycler reaches 95°C for initial denaturation, transfer the ARB2 reaction plate to the thermocycler.
7. Update the status of plates processed for the ARB2 reaction in the database if desired.

3.2.2. PCR Cleanup and Sequencing

1. Use a 12-channel pipette to transfer 5 µL of each ARB2 reaction to a new PCR reaction plate on ice.
2. Mix 2 µL EXOSAP-IT enzyme mix into the ARB2 reaction mix by pipetting up and down three times.
3. Seal plates with adhesive foil.
4. Move plate from ice to the thermocycler preheated to 37°C.
5. Incubate plates at 37°C for 15 min.
6. Incubate plates at 80°C for 15 min.
7. Update the status of plates subjected to PCR cleanup in the database if desired.
8. Add 13 µL freshly diluted sequencing primer at a concentration of 7.69 ng/µL to each sample for a final concentration of 5 ng/µL. For *MAR2xT7*, the Tn3 sequencing primer (**Fig. 1**) PMFLGM.GB-4a (5′-GACCGAGATAGGGTTGAGTG-3′) was used. Store samples at 4°C prior to sequencing.
9. Subject samples to fluorescently labeled dideoxynucleotide chain termination sequencing according to the kit manufacturer's instructions.

3.2.3. Uploading Sequences into the Relational Database and Sequence Analysis

Data-uploading methods are dependent on the database and software being used. Here we describe a general scheme that should be tailored accordingly.

1. Assign plate and sample names to sequencing data: Assemble ABI sequencing files into sets of 96, each set of 96 in a folder titled with a human-readable plate name that is recognizable by the database and linked to the unique identifier for the culture plate used as template for the sequences (**Note 16**). Each ABI file in the folder should have a human-readable and computer-parsable name that includes the culture plate name and the well position of the mutant from which the sequence was derived. ABI file names should be recognized by the database so that each ABI file is correctly associated with the unique

identifier for the proper mutant. For the PA14 library, each folder was given a three-digit numeric name that was linked to the unique culture plate database identifier. In the event that multiple sequencing attempts might be necessary to obtain data on all the mutants in a plate, the three-digit plate names were given a version number, each with its own unique database identifier. For example, folder 242v6 represents culture plate number 242, version 6. Sequences from individual wells in this plate are titled 242v6_A01, 242v6_A02, 242v6_A03, and so on.

3.2.4. Insertion Site Identification

Software used for data analysis should process sequence information in the following way to identify the insertion site of each mutant.

1. Assign a quality score to each base in the original sequence. Using the PHRED software application (www.phred.org), PATIMDB assigns a quality score for each base pair in each uploaded sequence or raw sequence.
2. Trim low-quality sequence. Bases with a quality score of less than 20 are trimmed off the raw sequence to produce a processed sequence (**Notes 17** and **18**).
3. Perform BLAST alignment with genomic sequences. PATIMDB aligns each processed sequence with both the PA14 and PAO1 genome sequences using the BLAST algorithm (***10***). The assignment of a location of the transposon insertion site in the PA14 genome for a given processed sequence isolated from a particular mutant is based on the region of the genome with the best BLAST score (**Note 19**).
4. Identify the transposon sequence immediately adjacent to the genomic sequence junction point. The precise location of the insertion site in the region of the BLAST hit is determined automatically using a Smith-Waterman algorithm built into PATIMDB that searches the first 120 bases of the raw sequence for alignment with a 26-base sequence at the end of the *MAR2xT7* transposon, allowing up to seven mismatches or gaps (**Fig. 2**). Parameters

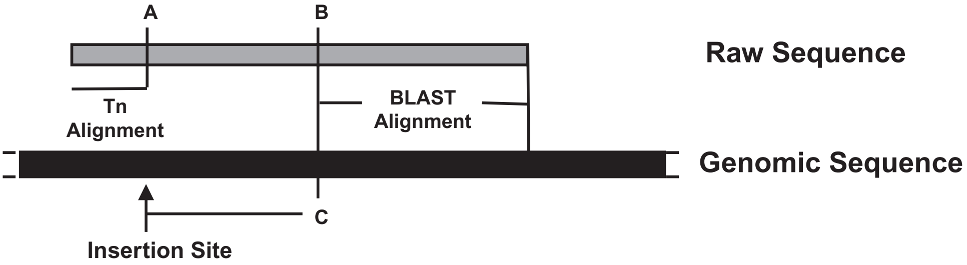


Fig. 2. Insertion site prediction methodology. PATIMDB (1) scans the raw sequence for the sequence aligning with the transposon sequence using the transposon sequence identification tool and (2) performs BLAST alignments with the genomic sequence. The raw sequence coordinates of the last base to align with the transposon sequence (*A*) and the first base to align with the genomic sequence (*B*) are used to determine the transposon insertion site in the genomic sequence. Insertion site (in genomic sequence coordinates) = *C*–(*B*–*A*), where *C* is the genomic sequence coordinate of the first aligning base. This rule applies regardless of which strand aligns with the raw sequence. In cases when the transposon sequence cannot be identified, A is given a fixed value based on where within the raw sequence the last base of a particular transposon is most frequently observed.

such as the maximum number of transposon sequence mismatches tolerated, how far 3′ the algorithm searches for the transposon junction, and how much of the transposon sequence is used in the search can be optimized for individual needs. Where the sequencing primer anneals with respect to the transposon junction, the sequence quality at the transposon junction and the accuracy of the zero positions of individual reads affect the accuracy of transposon sequence identification in the raw sequence. Setting these parameter requires trade-offs, however. For example, longer transposon junction search sequences require that a sequencing primer that anneals further upstream (in the 5′ direction) from the transposon junction be used, requiring a greater allowance for mismatches but allowing a higher degree of confidence that the determined transposon location is correct. For sequences in which the transposon sequence cannot be identified, a default insertion site is selected based on the observation that in most mutant sequences, the transposon junction point lies a set number of bases into the raw sequence (on average, 63 bp for *MAR2xT7*). Caution must be used, however. In several instances where no alignment with *MAR2xT7* was identified, the transposon sequence was found manually beyond the first 120 bases of the raw sequence, suggesting that the search window was set too narrowly. Using the default location of 63 bases in these cases puts the insertion site more than 63 bases away from the actual insertion site.

3.3. Library Mega-Analysis

3.3.1. Insert Distribution Across the Genome

To determine insertion coverage and to detect the presence of hot spots, the genomic coordinates of all transposon insertions are mapped into 1- or 10-kb bins. The number of mapped insertions in each bin is quantified. This analysis should be carried out routinely during library production to assess saturation of the genome.

3.3.2. Insert Distribution Within Predicted ORFs

Combining insertion site coordinates with the start and stop sites of every ORF in the genome gives the number of times each gene has been hit. This analysis also makes apparent which genes have not been hit while library production is in progress. If the fraction of undisrupted genes that are known to be nonessential in other organisms is high, library production should continue. Once near-saturation has been established, genes that were hit only once should be analyzed further to determine if there are more hits at the extreme 5′ and 3′ ends of these genes than would be predicted if insertions were completely random. An enrichment of hits at the extreme 5′ end of genes hit only once indicates possible transcriptional fusions with transposon-derived sequences. An enrichment of hits in the extreme 3′ end of genes hit only once suggests that the insertion did not disrupt gene function. Genes falling into either category may be essential.

3.3.3. Defining Essential Genes: Comparing Transposon Insertion Mutants in Two Different Strains

In the case of *P. aeruginosa*, transposon mutation libraries using different transposons have been constructed independently in strains PAO1 (*[11]* and **Chapter 9**) and PA14 *(1)*. Between the two libraries, the insertions sites of more than 60,000 transposon mutations have been mapped. By comparing the two libraries, we were able to identify orthologous genes that were not hit in either library and thereby determine a set of putative essential genes. First, we defined orthologs between the two strains (which we call PA14/PAO1 orthologs), and then the orthologs not hit in either library were compared. Genes not hit in either library were considered putative essential genes. As described in **Section 3.3.2**, for genes hit only one time, the hit distribution was skewed toward the extreme 5′ and 3′ ends of these genes, indicating that some PA14/PAO1 orthologs hit just once at either end of the coding sequence may be also putative essential genes (**Table 2**).

1. Identify orthologous genes in two different strains. Download and run the "findOrthologs.pl" program (http://ausubellab.mgh.harvard.edu/cgi-bin/pa14/downloads.cgi). This program generates a list of orthologs based on criteria such as the percentage identity and percentage difference of query length to alignment length. After the program is run, some manual curation of the list may be necessary. As a general rule, reciprocal best hits are selected as orthologs, while attempting to maintain synteny along the genome.
   The program requires:
   (a) FASTA format file or files containing the two single-contig genome sequences named "1" and "2."
   (b) FASTA format file or files containing predicted ORFs in both genomes. Each gene must occur only once and have a unique gene identifier (GeneID) in the title line.
   (c) Tab-separated values format file containing the fields "GeneID," "GenomeID," and "Start" indicating, respectively, the unique gene identifier, the genome identifier (1 or 2), and the start position of each gene.
   (d) Configuration file that contains cutoffs for percent sequence identity and the maximum difference in the length of individual BLAST query sequences.
2. Compare orthologous genes not hit in either library.
   (a) Generate a list of genes in each library that were not mutated (or not "hit") by identifying those genes from the total gene set in each genome. With a relational database, this requires a so-called left outer join in database parlance (**Note 20**).
   (b) Select one of the two mutated genomes as the reference genome. Perform a join of nonmutated genes from the nonreference genome library with the orthologs table (**Section 3.3.3, No. 1**) to generate a list of reference genome orthologs for these nonmutated genes.
   (c) Join the list of nonmutated genes in the reference genome with the list of reference genome orthologs of the nonmutated genes from the nonreference library (above). The

Table 2
Comparison of PA14 and PAO1 Orthologs Disrupted in Two Mutant Collections

	Predicted PA14/ PAO1 orthologs	PA14/PAO1 orthologs hit	PA14/PAO1 orthologs not hit	Unique insertion locations
P. aeruginosa strain PA14 (*1*)	5,102	3,954	1,148	22,881
P. aeruginosa strain PAO1 (*11*)	5,102	4,494	608	30,100
PA14/PAO1 orthologs not hit in either library			335	

intersection of these two sets gives a set of putative essential genes. This set excludes genes that are strain-specific or that lack identified orthologs.

Notes

1. Multiple matings over the course of library production also minimize the number of redundant mutants. Ten different matings were used to produce the PA14 library.
2. In control experiments related to the construction of the *P. aeruginosa* PA14 transposon mutation library, *MAR2xT7* transposants were first apparent at 2 h incubation, indicating that with the particular donor cells, transposon, and recipient cells used for these experiments, a 2-h mating time was sufficient to obtain a high frequency of transposition events but minimized the amplification of transposants, thereby reducing the frequency of isolating mutants containing the same mutation. It is important to determine the optimal time and donor:recipient culture ratio for matings with the particular strains to be used prior to scaling up.
3. The transposition frequency for a given mating combination must also be determined so that the number of mating spots that must be harvested to obtain the desired number of transposants is known. For PA14 mated with MC4100/p*MAR2xT7* and HB101/pRK2013, three pooled mating spots consistently generated more than 4800 transposants. Transposant colonies approximately 2.5 mm in width can be recognized by the QBot.
4. The antibiotic resistance marker on the backbone of the donor plasmid carrying the transposon serves to verify that only the transposon has inserted into the genome, and integration of the entire donor plasmid has not occurred. In initial experiments, putative transposants should be tested for donor plasmid integration.
5. In the case of *P. aeruginosa* strain PA14, we have observed that liquid cultures grown statically under microaerobic conditions contain a high frequency of so-called phenotypic variants that are resistant to high concentrations of multiple antibiotics *(7, 8)*. These antibiotic-resistant variants (called RSCVs for **r**ough **s**mall **c**olony **v**ariants) exhibit a variety of transient phenotypic changes in addition to antibiotic resistance including high surface hydrophobicity that results in increased biofilm formation and reduced virulence. We do not know whether RSCV formation is a common feature of all *P. aeruginosa* strains or other bacterial species. Because the frequency of RSCVs increases with high levels of antibiotics, we determined the minimal concentrations of gentamicin and Irgasan required to select for PA14/*MAR2xT7* transposants. Gentamicin resistance is encoded by *MAR2xT7*, and PA14 is naturally resistant to Irgasan. Therefore, Irgasan is used to select for *P. aeruginosa* and against *E. coli*.
6. If a QBot is not available, colonies can be picked by hand, but depending on the library size, this could be a daunting task.
7. It is essential to set the QBot picking parameters to ensure that the robot does not mistake two overlapping colonies for a single colony. Look at the trays after the picking run to verify that only individual colonies were disrupted by the pins.
8. Trays can be stored at 4°C and repicked if necessary.
9. As previously described, experiments in our lab with *P. aeruginosa* strain PA14 showed that static long-term culture (more than 16 h) greatly increases the frequency of RSCVs in PA14 cultures. However, growing PA14 in deep-well Costar/Corning blocks (600 µL media in 2.0 mL 96-well deep-well blocks) with agitation in a HiGro shaker (**Materials**) or in a standard plate shaker that can hold blocks for 16 h or less prevents RSCV formation and makes the cultures easier to transfer (the cultures tend to have a more uniform consistency).

However, because the QBot can only inoculate low-profile plates, inoculation of deep-well blocks had to be done by hand.

10. In the construction of the PA14 library, an individual rack of transfer tips was used three times to transfer the same cultures in **steps 4, 10, 11,** and **12**. Moreover, to keep costs down, the tips used to add glycerol to culture plates by dispensing glycerol from above the plate into the cultures were saved and reused for 24 different culture plates. Subsequently, however, we found by carrying out quality-control experiments that the tips used to dispense glycerol did become contaminated with PA14. We therefore recommend instead that fresh tips be used to add glycerol to each plate.
11. This example illustrates the importance of quality-control testing to ensure that the liquid-handling method used to aliquot cultures be cross-contaminant free. Using a control plate consisting of some wells inoculated with the mutagenized strain interspersed with many uninoculated (sterile) wells is a simple way to determine the level of potential cross-contamination at each step of the protocol. Store source and destination plates at 37°C for several days to confirm the absence of cell growth in uninoculated wells. Control plate tests of our methods defined several critical parameters essential for minimizing cross-contamination when handling PA14 cultures. These parameters are discussed below.
12. The following guidelines arose from thorough testing of the Biomek FX method used to transfer PA14 cultures.
    (a) *Glycerol (to final concentration of 15%) must be added and mixed into PA14NR set cultures before transfer.* Transferring PA14 cultures grown in LB or LB + 15 µg/mL gentamicin for various lengths of time, with or without agitation, to either 96- or 384-well plates, resulted in a high frequency of cross-contamination of both adjacent and nonadjacent wells. This occurred whether the transfer was performed by the Biomek FX robot or by hand using a multichannel pipettor. We assume this cross-contamination is the result of aerosols from the tips as they are held over destination plates. We found that the addition of glycerol (final concentration of 15%) prior to transfer greatly decreased the frequency of cross-contamination of wells. Whatever the cause of cross-contamination, it is essential that 15% glycerol be added to cultures to be transferred when making copies of the library. Because glycerol inhibits the PCR reaction, transfer of culture to be used as PCR templates from culture plates should be performed before addition of glycerol. This step, therefore, must be thoroughly tested for well-well cross-contamination.

    (b) *Avoid carryover of culture mix on transfer tips.* Culture aspiration and dispensing is prone to drops of culture hanging from tip ends that can easily cross-contaminate wells as the robotic head moves over the plates laid out on the deck. To avoid this, we programmed the Biomek FX to touch the tips to the side of the wells with each aspiration and dispensing step (*see* below). In addition, all "blow-out" steps were skipped because, in our hands, this formed bubbles of culture/glycerol mix on the ends of the tips, a potential source of contamination as the tips move over the blocks and plates on the deck.

    (c) *Library propagation should be performed in a 96-well format.* Even with glycerol addition to cultures prior to transfer, we were unable to inoculate 384-well plates either by hand or robotically without cross-contaminating adjacent wells. This is presumably due to the necessity of touching the tips to the side of the wells both after aspiration of culture in source plates and after dispensing culture in destination plates. The mostly likely reason why the wells in 384-well plates get cross-contaminated is that the well walls are shared between wells. In contrast, well walls in 96-well plates are not shared between adjacent wells.

(d) *Keep culture times to a minimum.* We generally grew *P. aeruginosa* PA14 cultures for 14h but not more than 16h. Although some mutants may take longer to get to saturation, they are almost always at a high enough density after 14h to be used for transfer to copy plates.

(e) *Grow cultures with agitation in deep-well blocks.* Reduced aeration of PA14 cultures greatly increases the incidence of RSCV formation. We have tested cultures grown in a Genomic Solutions, Inc., HiGro shaker/incubator with supplemental O_2 and cultures grown in a regular shaker with no extra O_2 added. Both conditions prevented the appearance of variants.

(f) *Keep the surface-to-volume ratio high, ensuring proper aeration of the culture.* Growing cultures larger than 750µL in 2.0-mL 96-well culture blocks resulted in the appearance of variants. When growing large cultures, we grew 600-µL cultures in deep-well blocks (catalog no. 7556–9600, USA Scientific, Ocala FL), resulting in a surface-to-volume ratio of 0.92.

(g) *Avoid extensive dilution.* The larger the degree of dilution, the higher the frequency of variants. We assume this is because the increased amount of cell division required to saturate the culture promotes the appearance of variants in the culture. Preliminary tests have shown that inoculating 600µL media with an average-size wild-type colony picked with a pipette tip is sufficient to prevent the appearance of variants in the saturated culture. When using thawed liquid stocks to inoculate media, we generally diluted 1:50.

(h) *Keep surface area of tip coated with culture/glycerol mix to a minimum.* Even when source cultures contain 15% glycerol, transfer of cultures will result in well-to-well cross-contamination of source and destination plates if the majority of the surface area of the transfer tip is coated with culture/glycerol mix. To avoid this problem, the Biomek FX is programmed to have transfer tips aspirate culture mix no more than 10mm below the liquid surface.

(i) *Seal plates on the deck of the robot.* The simple act of moving the plates from the robot to the bench top before sealing has led to cross-contamination in test runs. Seal the plates thoroughly on the deck, using a roller to ensure that each well is firmly sealed.

13. QFilled plates can be stored at 4°C overnight before run.
14. Others have modified our protocol, replacing DMSO with 1.25M Betaine (catalog no. 14300; Fluka, Sigma Aldrich, St. Louis, MO) with excellent sequencing success rates (D. Ewen Cameron and J. Mekalanos, personal communication).
15. Originally, ARB1 PCR was performed using the ARB1 primer (**Table 1**), with a success rate of approximately 95%. Over time, the efficiency of sequencing with this primer dropped. We found that many of the products of sequencing were extremely short—only as long as the transposon sequence—suggesting that under the conditions of the particular PCR reaction employed, the ARB1 primer had an affinity for the end of the transposon. Several additional ARB primers were created (**Table 1**), including two (ARB1D and ARB1E) in which the defined sequence was changed to be less likely to hybridize to the transposon sequence. ARB1D and ARB1E gave the greatest sequencing success rates (~75%).
16. When using PATIMDB to analyze sequences, each plate folder containing 96 sequences is in turn placed into a folder entitled with the transposon name (2xT7). The transposon folder is placed in a folder titled “PA14.” PATIMDB is launched and “Sequence Analysis” is selected. Each folder is uploaded into PATIMDB individually.
17. We found that a Phred quality score of 20 adequately filtered out poor sequences; however, different values can be used if desired.

18. Because the sequence quality at the very beginning of reads is often poor, it is advisable to have the sequencing primer positioned a sufficient number of base pairs from the transposon junction so that at least 30 bases of the transposon sequence are routinely present in the resulting sequencing reads beyond the region of poor sequence.
19. In most cases for the PA14 library, alignment with one region in the genome had a high BLAST score, whereas other genomic regions had much lower BLAST scores. The difference between the best BLAST score and the second-best BLAST score, which we called the Bit Score Separation, was large, indicating a high confidence that the region with the highest BLAST score was the site of the insertion. Therefore, the position of the insertion was based on that BLAST hit. If, however, more than one region of the genome aligns well with a raw sequence and the Bit Score Separation is zero (i.e., in cases of gene duplications), the insertion location cannot be unambiguously determined.
20. If a relational database is not available, a spreadsheet can be used to compare lists of non-mutated orthologs in each library. Orthologs that are common between the two lists are putative essential genes.

Acknowledgments

Funding was provided by the NIH Heart Lung and Blood Institute HL66678, NIH R01, AI064332, and a grant from the Cystic Fibrosis Foundation.

References

1. Liberati, N. T., Urbach, J. M., Miyata, S., Lee, D. G., Drenkard, E., Wu, G., et al. (2006) An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2833–2838.
2. Lampe, D. J., Grant, T. E., and Robertson, H. M. (1998) Factors affecting transposition of the *Himar1 mariner* transposon *in vitro*. *Genetics* **149**, 179–187.
3. Rubin, E. J., Akerley, B. J., Novik, V. N., Lampe, D. J., Husson, R. N., and Mekalanos, J. J. (1999) *In vivo* transposition of *mariner*-based elements in enteric bacteria and mycobacteria. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 1645–1650.
4. Figurski, D. H., and Helinski, D. R. (1979) Replication of an origin-containing derivative of plasmid RK2 dependent on a plasmid function provided *in trans*. *Proc. Natl. Acad. Sci. U.S.A.* **76**, 1648–1652.
5. Adamczyk, M., and Jagura-Burdzy, G. (2003) Spread and survival of promiscuous IncP-1 plasmids. *Acta Biochim. Pol.* **50**, 425–453.
6. Jones, D. L., and King, S. (1954) Evaluation of a single medium for detecting production of urease and indole. *Am. J. Clin. Pathol.* **24**, 1316–1317.
7. Drenkard, E., and Ausubel, F. M. (2002) *Pseudomonas* biofilm formation and antibiotic resistance are linked to phenotypic variation. *Nature* **416**, 740–743.
8. Drenkard, E. (2003) Antimicrobial resistance of *Pseudomonas aeruginosa* biofilms. *Microbes Infect.* **5**, 1213–1219.
9. Caetano-Anolles, G., and Bassam, B. J. (1993) DNA amplification fingerprinting using arbitrary oligonucleotide primers. *Appl. Biochem. Biotechnol.* **42**, 189–200.
10. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
11. Jacobs, M. A., Alwood, A., Thaipisuttikul, I., Spencer, D., Haugen, E., Ernst, S., et al. (2003) Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 14339–14344.

11

The Construction of Systematic In-Frame, Single-Gene Knockout Mutant Collection in *Escherichia coli* K-12

Tomoya Baba and Hirotada Mori

Summary

Here we describe the systematic construction of well-defined, in-frame, single-gene deletions of all nonessential genes in *Escherichia coli* K-12. The principal strategy is based on the method for one-step inactivation of chromosomal genes in *E. coli* K-12 established by Datsenko and Wanner (*1*), namely, the replacement of a target gene with a selectable antibiotic-resistant marker generated by polymerase chain reaction (PCR) using oligonucleotide DNA primers homologous to the gene flanking regions. The advantages of this method include complete deletion of an entire open reading frame and precise design eliminating polar effects for the downstream genes on *E. coli* chromosome.

Key Words: complete deletion; *Escherichia coli*; FLP recombinase; FRT; gene knockout; homologous recombination; in-frame deletion; lambda Red recombinase; mutant; site-specific recombination.

1. Introduction

A single-gene–deleted mutant collection should provide a fundamental tool for "reverse genetics" approaches, permitting analysis of the consequences of the complete loss of gene function, in contrast with forward genetics approaches, in which mutant phenotypes are associated with the corresponding genes. In the *Saccharomyces cerevisiae* functional genomics project, a nearly complete set of single-gene deletions covering 96% of yeast annotated open reading frames (ORFs) was constructed by using a polymerase chain reaction (PCR) gene replacement method (*2*). The yeast mutants were isolated by direct transformation with PCR products encoding kanamycin resistance and containing 45-nt flanking homologous sequences for adjacent chromosomal regions. Genome-scale disruption of *Bacillus subtilis* genes (*3*) was done by inactivating each gene with a gene-specific plasmid clone. Comprehensive transposon mutagenesis of *Pseudomonas aeruginosa* was carried out by generating a large set (~30,100) of sequence-defined mutants (*4*). The construction of an *Escherichia coli* gene disruption bank was initiated by using transposon mutagenesis (*5*). The strategy was to mutagenize the *E. coli* chromosomal regions carried in each Kohara phage clone (*6*) and then to

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

recombine these mutations onto the host chromosome by homologous recombination (**Chapter 13**). Although this approach yielded a large, unique collection of mutants, the methodology was laborious. First, it was necessary to determine the transposon insertion site. Second, complications resulting from transposon mutagenesis, such as incomplete disruption of the targeted gene and polar effects on the downstream genes, were unavoidable. Importantly, during the course of our attempts to create an *E. coli* mutant bank by transposon mutagenesis, Datsenko and Wanner reported a novel, highly efficient method of gene disruption using the phage lambda Red recombination system (*1*). The strategy was analogous to the PCR-based gene-deletion method utilized in yeast, except *E. coli* cells were carrying a low-copy-number, replication-thermosensitive (*ts*-ori) plasmid (easily curable at 30°C) for expressing the lambda Red recombinase (**Fig. 1**) (*1*). Advantages of this method include the ability to target *E. coli* genes for complete deletion, the ability to design the deleted region arbitrarily and precisely, and the ability to easily eliminate the antibiotic-resistance gene if necessary (*1*). Here, we describe the systematic construction of well-defined, single-gene deletion mutants

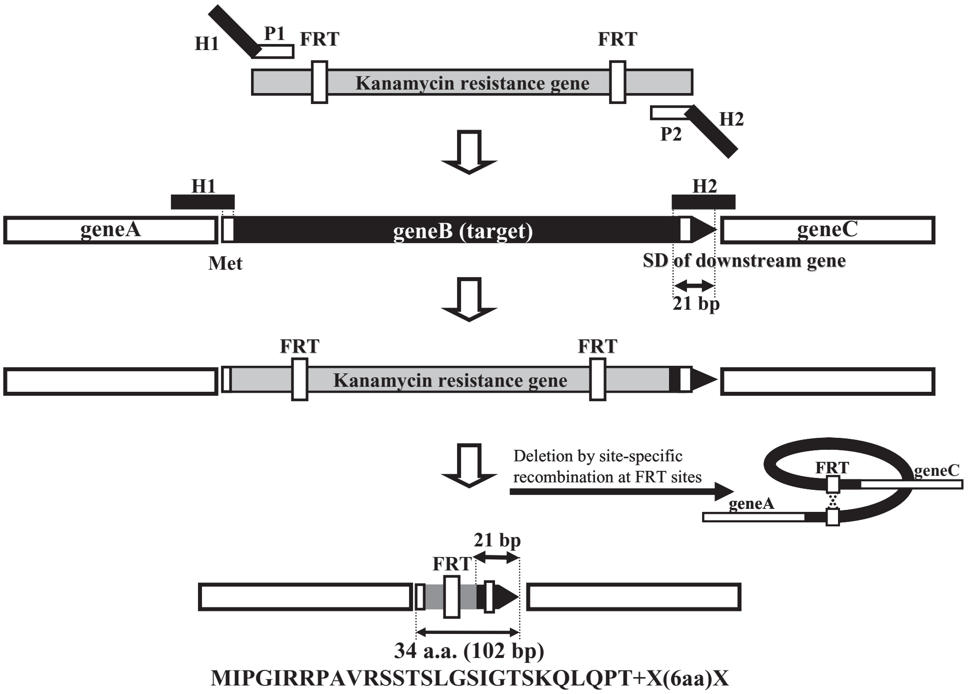

H1
P1
FRT
FRT
Kanamycin resistance gene
P2
H2
H1
H2
geneA
geneB (target)
geneC
Met
SD of downstream gene
21 bp
FRT
FRT
Kanamycin resistance gene
Deletion by site-specific recombination at FRT sites
FRT
geneC
geneA
21 bp
FRT
34 a.a. (102 bp)
MIPGIRRPAVRSSTSLGSIGTSKQLQPT+X(6aa)X

Fig. 1. Primer design and construction of single-gene deletion mutants. Gene knockout primers have 20-nt 3′ ends for priming upstream (P1) and downstream (P2) of the FRT sites flanking the kanamycin-resistance gene in pKD13 and 50-nt 5′ ends homologous to upstream (H1) and downstream (H2) chromosomal sequences for targeting the gene deletion (*8*). H1 includes the gene B (target) initiation codon. H2 includes six C-terminal codons, the stop codon, plus 29-nt downstream. The same primer design with respect to gene B was used to target deletions regardless of whether gene B lies in an operon with genes A and C, as shown, or in any different chromosomal arrangements. SD, Shine-Dalgarno ribosome binding sequence.

that are designed to yield in-frame deletions upon excision of the resistance gene to eliminate polar effects on downstream genes.

2. Materials

1. *E. coli* strains BW25113 [*rrnB3 ΔlacZ4787 hsdR514 Δ(araBAD)567 Δ(rhaBAD)568 rph-1*] and BW25141 [*rrnB3 ΔlacZ4787 ΔphoBR580 hsdR514 Δ(araBAD)567 Δ(rhaBAD)568 galU95 ΔendA9::FRT ΔuidA3::pir(wt) recA1 rph-1*] (*1*).
2. Plasmids pKD13 (GenBank accession no. AY048744) and pKD46 (GenBank accession no. AY048746) (*1*).
3. Bacto Tryptone (BD, Franklin Lakes, NJ).
4. Yeast extract (BD).
5. NaCl (Wako, Osaka, Japan).
6. Luria-Bertani (LB) medium (*7*).
7. Antibiotics ampicillin (Wako) and kanamycin (Wako).
8. Glucose (Wako).
9. L-Arabinose (Wako).
10. *Dpn*I restriction enzyme (New England Biolabs, Beverly, MA).
11. *TaKaRa Ex Taq* polymerase (Takara Shuzo Inc., Kyoto, Japan).
12. SeaKem GTG Agarose (Takara Shuzo Inc.).
13. Oligonucleotide DNA primers (Nihon Idenshi Inc., Sendai, Japan).
14. 2.5 mM each of deoxynucleotide triphosphate (dNTP) mixture (Takara Shuzo Inc.).
15. 0.2-cm electroporation cuvette (Bio-Rad, Hercules, CA).
16. SOC medium: 2% Bacto Tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCL, 10 mM $MgCl^2$, 10 mM $MgSO^4$, 20 mM glucose (*7*).
17. Bromophenol Blue (BPB; Wako).
18. Glycerol (Wako).
19. E-Gel 96 system (Invitrogen, Carlsbad, CA).
20. Tris-Acetate (TAE) electrophoresis running buffer: 40 mM Tris-acetate, 1 mM EDTA, pH 8.3.
21. 6× gel loading dye solution: 1 mM EDTA, 30% glycerol, 1.5 mg/mL BPB in Milli-Q water
22. 50-mL sterilized plastic tube (Nunc, Rochester, NY).
23. 15-mL sterilized plastic tube (Nunc).
24. 96-well microtiter plate (Nunc).
25. 96-well PCR reaction plate (Applied Biosystems, Foster City, CA).
26. Aluminum seal for 96-well plate (Applied Biosystems).
27. 96-well full plate cover (Applied Biosystems).

3. Methods

The methods described below outline (1) the primer design for the PCR fragments for in-frame deletions, (2) the amplification and purification of PCR fragments, (3) the preparation of *E. coli* electroporation-competent cells, (4) the *E. coli* transformation with PCR fragments, and (5) the verification of gene knockout mutants.

3.1. Primer Design

The primer design for PCR amplification of DNA constructs for in-frame deletions is described in **Sections 3.1.1** and **3.1.2**. This includes (1) the description of the pKD13 marker-DNA template vector and (2) the description of the basic structure of PCR primers.

3.1.1. pKD13 Marker-DNA Template Vector

The plasmid pKD13 was specifically constructed as a marker-DNA template vector for gene disruption *(1)*. FRT (FLP recognition target) sites were adjacent on both sides of the kanamycin-resistant gene cassette (**Fig. 1**).

3.1.2. PCR Primers

PCR primers for constructing gene deletions included 50-nt homologous to the adjacent upstream or downstream flanking regions of the target gene and 20-nt 3′ end for amplification of kanamycin (*kan*) resistance gene and the nearby FRT sites in pKD13. N-terminal deletion primers had a 50-nt-long 5′ extension including the gene initiation codon (H1) and the 20-nt sequence 5′-ATTCCGGGGATCCGTCGACC-3′ (P1). C-terminal deletion primers consisted of 21-nt for the C-terminal region, the termination codon, and 29-nt downstream (H2) and the 20-nt sequence 5′-TGTAGGCT GGAGCTGCTTCG-3′ (P2; **Fig. 1**). The targeting PCR products were designed to create in-frame deletions. In this case, a targeted ORF was deleted almost entirely (from the second through the seventh codon from the C-terminus), leaving the start codon and translational signal for a downstream gene intact [**Fig. 1** *(8)*; *see* **Note 1**].

3.2. Amplification and Purification of PCR Fragments

The PCR fragments for gene deletion were amplified and purified in 96-well microplates, described in **Sections 3.2.1** and **3.2.2**. This includes (1) the description of PCR fragment amplification from pKD13 and (2) the description of purification of the PCR fragments (**Note 2**).

3.2.2. PCR Fragment Amplification

PCR reactions were done in 50-μL reactions containing 2.5 U of *TaKaRa Ex Taq* polymerase, 1 pg pKD13 DNA, 1.0 μM of each primer, and 200 μM dNTPs (**Note 3**). Reactions were run for 30 cycles: 94°C for 30 s, 59°C for 30 s, and 72°C for 2 min plus an additional 2 min at 72°C.

3.2.3. PCR Fragment Purification

PCR products were digested with *Dpn*I and ethanol-precipitated to purify the PCR products from contaminating template plasmid DNA and excess primers. Finally, they were resuspended in 6 μL H_2O, and 1 μL of each sample was analyzed by 1% agarose gel electrophoresis using 0.5× Tris-Acetate (TAE) buffer or the E-Gel 96 system (**Note 4**). All procedures are carried out in 96-well formatted PCR reaction plates, and the details are described below.

3.2.3.1. Ethanol Precipitation in 96-Well Format

1. Add 120 μL of 100% ethanol into PCR-amplified DNA solution (48 to 50 μL) by multichannel pipette or appropriate dispensing robot systems and mix by pipetting (**Note 5**).
2. Centrifuge with a 96-well full plate cover (3800 ×*g*; 30 min, 20°C) (**Note 6**).

3. Remove ethanol by putting 96-well plates upside down on a Kim-towel.
4. Add 200 μL of 70% ethanol (**Note 7**).
5. Centrifuge with a 96-well full plate cover (3800 ×*g*; 20 min, 20°C) (**Note 6**).
6. Remove ethanol by putting 96-well plates upside down on Kim-towel.
7. Remove ethanol completely by brief centrifugation (**Note 8**), up to 8 Xg and placing the plates upside down on Kim-towel.
8. Dry completely in a PCR machine, setting it at 55°C for 5 to 10 min; keep the plate covered with a Kim-wipe.

3.2.3.2. *DpnI* Treatment

1. Prepare *Dpn*I reaction mix containing (per well):

10× buffer	3 μL (NEB buffer 4)
*Dpn*I	0.3 μL (6 units)
ddH_2O	27 μL

2. Dispense 30 μL of *Dpn*I reaction mix into each well and mix by pipetting.
3. Flush by centrifugation.
4. Incubate at 37°C for 1.5 h covering PCR plate with a rubber cap.
5. Ethanol-precipitate as described above.
6. Dissolve in 6 μL ddH_2O.
7. Check 1 μL of each sample by agarose gel electrophoresis, using for example the E-Gel 96 System (**Note 9**).
8. Store the rest (5-μL DNA solutions) at –20°C for *E. coli* transformation.

3.3. Preparation of E. coli *Electroporation-Competent Cells*

The preparation of *E. coli* K-12 BW25113 electroporation-competent cells carrying the Red helper plasmid pKD46 is performed largely according to the methods described in Refs. *1* and *7*; however, slightly modified procedure for a large-scale preparation is described below.

1. Incubate the preculture in 50 mL SOB medium with ampicillin (50 μg/mL) overnight at 30°C with vigorous aeration.
2. Inoculate 400 mL SOB medium with 2 mM L-arabinose in 3-L flasks with 4 mL of the overnight preculture. Prepare six flasks (**Note 10**) and incubate the flasks at 30°C with agitation.
3. Measure the OD_{600} of growing culture. When it reaches 0.3, rapidly transfer the flask to an ice-water bath for 15 min.
4. Transfer the cultures to ice-cold centrifuge bottles. Harvest the cells by centrifugation at 1500 ×*g* for 10 min at 4°C. Decant the supernatant and resuspend the cell pellet in 300 mL of ice-cold pure H_2O (**Note 11**).
5. Harvest the cells by centrifugation at 1500 × *g* for 10 min at 4°C. Decant the supernatant and resuspend the cell pellet in 150 mL ice-cold pure H_2O (**Note 11**).
6. Harvest the cells by centrifugation at 1500 × *g* for 10 min at 4°C. Decant the supernatant and resuspend the cell pellet in 90 mL ice-cold 10% glycerol (**Note 11**).
7. Harvest the cells by centrifugation at 1500 × *g* for 10 min at 4°C. Decant the supernatant and resuspend the cell pellet in 40 mL ice-cold 10% glycerol (**Note 11**).
8. Harvest the cells by centrifugation at 1500 × *g* for 10 min at 4°C. Decant the supernatant and resuspend the cell pellet in 8 mL ice-cold 10% glycerol (**Note 11**).

9. Harvest the cells by centrifugation at 1500 × *g* for 10 min at 4°C. Carefully decant the supernatant and use a pipette to remove any remaining drops (**Note 12**). Resuspend the cell pellet in 1.6 mL ice-cold 10% glycerol.
10. For storage, dispense 50-µL aliquots of the cell suspension into sterile, ice-cold 1.5-mL tubes and transfer to a −70°C freezer (**Note 13**).

3.4. E. coli *Transformation with PCR Fragments*

3.4.1. Electroporation

Fifty microliters of competent cells are mixed with 400 ng of the PCR fragment in an ice-cold 0.2-cm cuvette. Cells are electroporated at 2.5 kV with 25 µF and 200 ohm, immediately followed by addition of 1 mL of SOC medium with 1 mM L-arabinose. After incubation for 2 h at 37°C, one-tenth portion was spread onto LB agar plate to select for KmR transformants at 37°C (**Notes 14** and **15**).

3.4.2. Storage of Gene Knockout Mutants

Eight independent colonies are transferred into 150 µL LB medium with kanamycin in 96-well microplates and incubated overnight at 37°C without shaking. After growth check, sterile glycerol is added to final concentration of 15%, each 96-well microplate is sealed with an aluminum seal and stored at −80°C (**Fig. 2**).

3.5. Verification of Gene Knockout Mutants

From every gene deletion experiment, four or eight Km^R colonies were chosen and checked for ones with the correct structure by PCR using a combination of locus- and kanamycin-specific primers (**Fig. 3**; *see* **Note 16**). Mutants were scored as correct if two or more colonies had the expected structure based on PCR tests for both upstream and downstream junctions *(8)*.

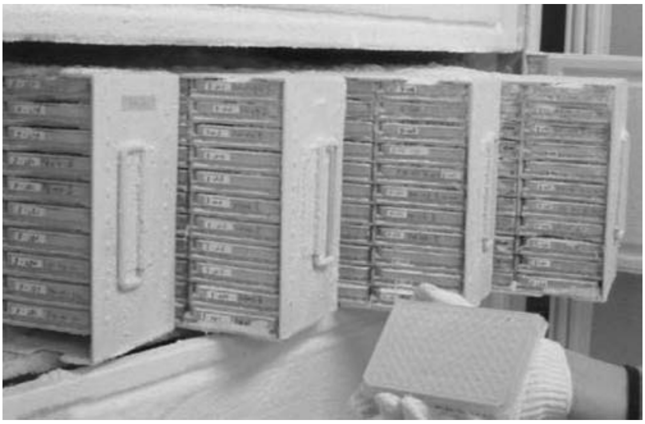


Fig. 2. Storage of gene knockout mutants in the −80°C deep freezer. Eight independent colonies from each gene knockout experiment were transferred into 150-µL LB medium with kanamycin in 96-well microplates and incubated overnight at 37°C without shaking. Mutants were stored in 15% glycerol at −80°C in 96-well microplates.

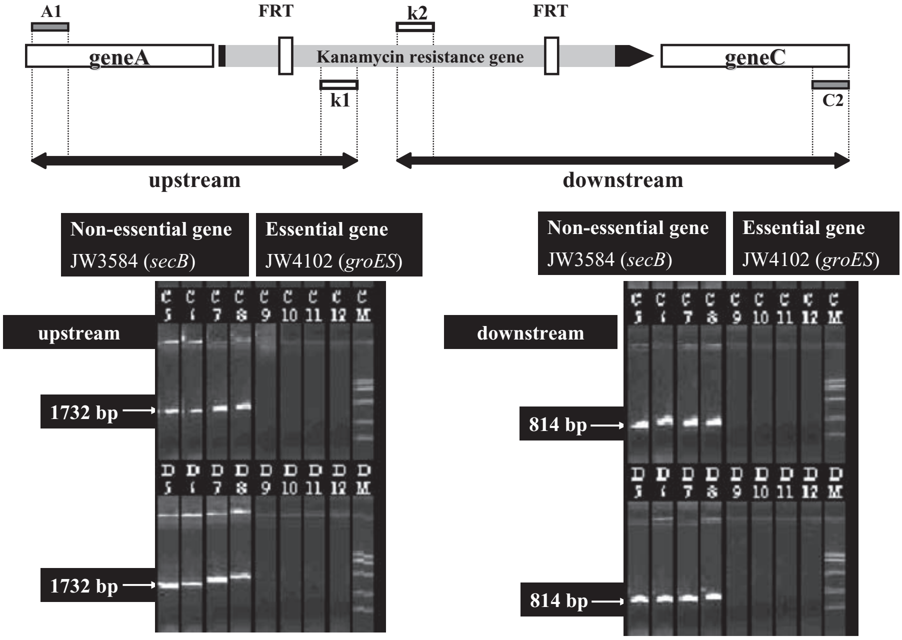


Fig. 3. PCR verification of deletion mutants. Two PCR reactions were carried out to confirm correct genome structure of all deletion mutants with two primers (k1 and k_2 in *kan*) and another two primers (A1 and C2) for upstream and downstream genes of the targeted gene, respectively. The two PCR reactions to confirm the upstream (A1–k1) and downstream (k2–C2) structures, respectively. Mutants were scored as correct if two or more colonies had the expected structure based on PCR tests for both junction fragments.

PCR tests were performed in 10-µL reactions containing 0.5 U *TaKaRa Ex Taq* polymerase, 0.5 µM of each primer, and 250 µM dNTPs. Reactions were "hot started" at 95°C for 2 min and run for 30 cycles: 94°C for 30 s, 60°C for 30 s, 72°C for 5 min, plus an additional 2 min at 72°C (**Note 17**).

Notes

1. Chromosomal genes were targeted for mutagenesis with PCR products containing a resistance cassette flanked by FRT sites and 50-bp homologies to adjacent chromosomal sequences (**Fig. 1**). To reduce polar effects on the downstream gene expression, primers were designed so that excision of the resistance cassette with the FLP recombinase would create an in-frame deletion of the respective chromosomal gene (**Fig. 4**). Primer sequences were based on the highly accurate *E. coli* K-12 genome sequence ***(9)*** in which the majority of the corrections to coding regions and start codon reassignments had been made in accordance with the November 2003 *E. coli* K-12 annotation workshop ***(10)***.
2. All experiments were performed in the 96-well format for higher throughput and reliability.
3. The amount of the template pKD13 plasmid DNA in a PCR reaction mixture should be minimal (i.e., less than 1 pg). This facilitates the complete removal of pKD13 by *Dpn*I treatment, an essential purification step required for reduction of background clones during the subsequent *E. coli* transformation with the PCR products.

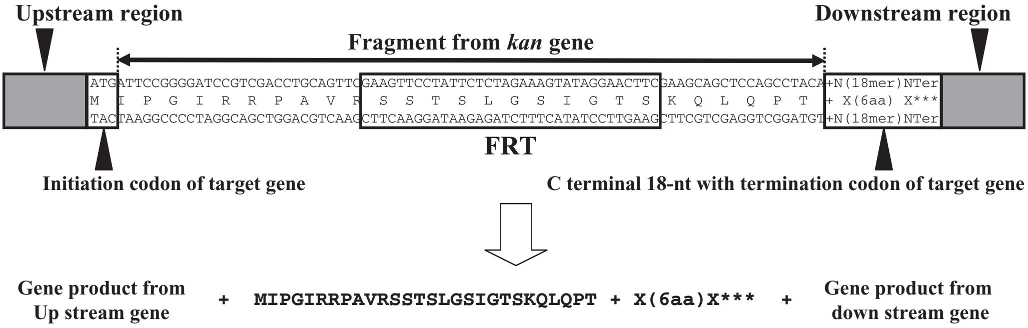

Upstream region
Downstream region
Fragment from *kan* gene
ATGATTCCGGGGATCCGTCGACCTGCAGTTCGAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCGAAGCAGCTCCAGCCTACA+N(18mer)NTer
M I P G I R R P A V R S S T S L G S I G T S K Q L Q P T + X(6aa) X***
TACTAAGGCCCCTAGGCAGCTGGACGTCAAGCTTCAAGGATAAGAGATCTTTCATATCCTTGAAGCTTCGTCGAGGTCGGATGT+N(18mer)NTer
FRT
Initiation codon of target gene
C terminal 18-nt with termination codon of target gene
Gene product from Up stream gene
+ MIPGIRRPAVRSSTSLGSIGTSKQLQPT + X(6aa)X*** +
Gene product from down stream gene

Fig. 4. Structure of in-frame deletions. FLP-mediated excision of the FRT-flanked resistance gene is predicted to create a translatable scar sequence in-frame with the gene B target initiation codon and its C-terminal 18-nt coding region. Translation from the authentic gene B SD (Shine-Delgarno ribosome binding sequence) and start codon is expected to produce a 34-residue scar peptide with an N-terminal Met, 27 scar-specific residues, and 6 C-terminal, gene B–specific residues.

4. For the high-throughput electrophoretic analysis of PCR products, the E-Gel 96 System (Invitrogen) is very useful because the 96-well-formatted electrophoresis corresponds with the format of 96-well plates with PCR reaction mixtures (**Fig. 5**).
5. Add 100% salt-free ethanol because PCR reaction buffer contains enough salts for efficient DNA precipitation.
6. The centrifugation conditions given are for Beckman R25–type centrifuge.
7. Never mix by pipetting because precipitated DNA pellets are easily released from the walls of a 96-well PCR plate.
8. To avoid dislodging the pellets, centrifugate the plates very briefly (at 8 × *g* for 1 s) with slowest possible acceleration.
9. Add 19 µL of 1× loading dye solution to 1 µL of each DNA solution, mix by pipetting, and load onto E-gel 96 agarose gel.
10. The volume of culture and number of flasks depend on the shaker and centrifuge facilities.
11. Gentle swirling is better than vortexing.
12. Be careful when decanting because the cell pellets lose adherence in 10% glycerol.
13. Ideally, in this protocol 192 1.5-mL tubes will be prepared as competent cells.
14. In our protocol, one "transformation experimental unit" consisted of 24 targeted genes. Incubation at 37°C was followed by a series of eight genes' electroporations. The four series of eight shocks took almost 1 h. Complete processing of an entire "transformation unit" took about 2 h, including at least 1 h incubation at 37°C and enough time for spreading the cells on LB agar plates after electroporation.
15. Our standard protocol usually yielded 10 to 1000 Km^R colonies when cells were incubated aerobically at 37°C on LB agar containing 30 µg per mL kanamycin. The most critical step was the preparation of highly electrocompetent cells (>10^9 transformants per 1 µg of plasmid DNA under standard conditions). Mutants were isolated in batches, and each batch included a positive control (PCR product for disruption of *ydhQ*) and a negative control (no PCR product added) samples. The latter usually yielded only 10 to 100 tiny background colonies (**Fig. 6**).
16. The verification of gene knockout mutants by genomic PCR depends on the Tm of gene-specific primers (**Fig. 3**), which were used in the construction of the ASKA library ***(11)***, local genomic structure, and so on.

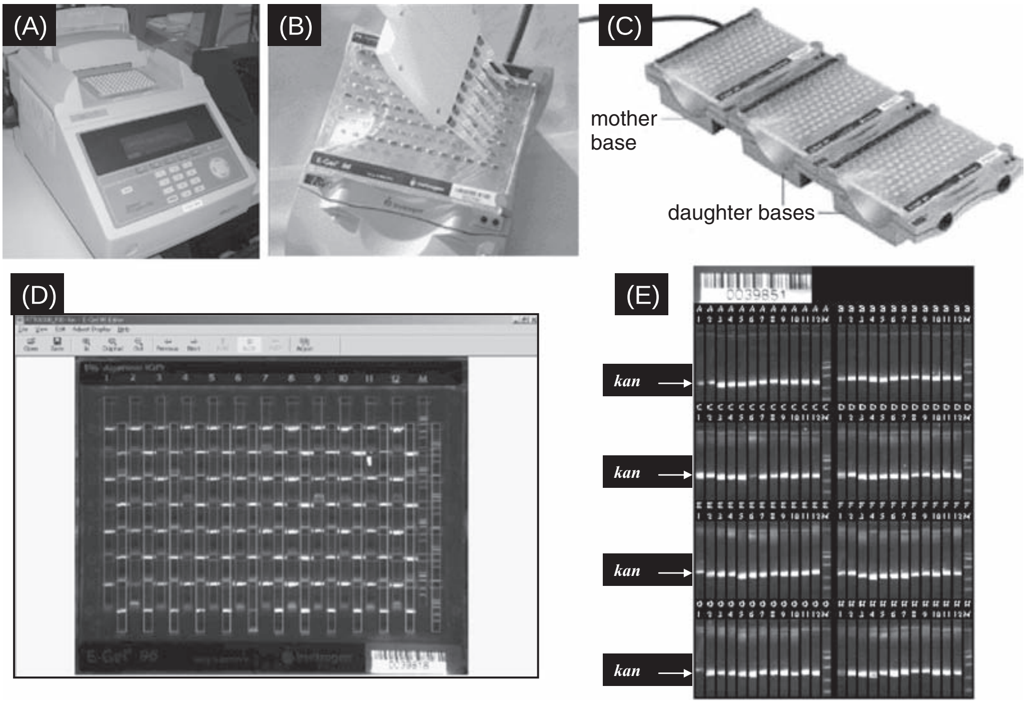
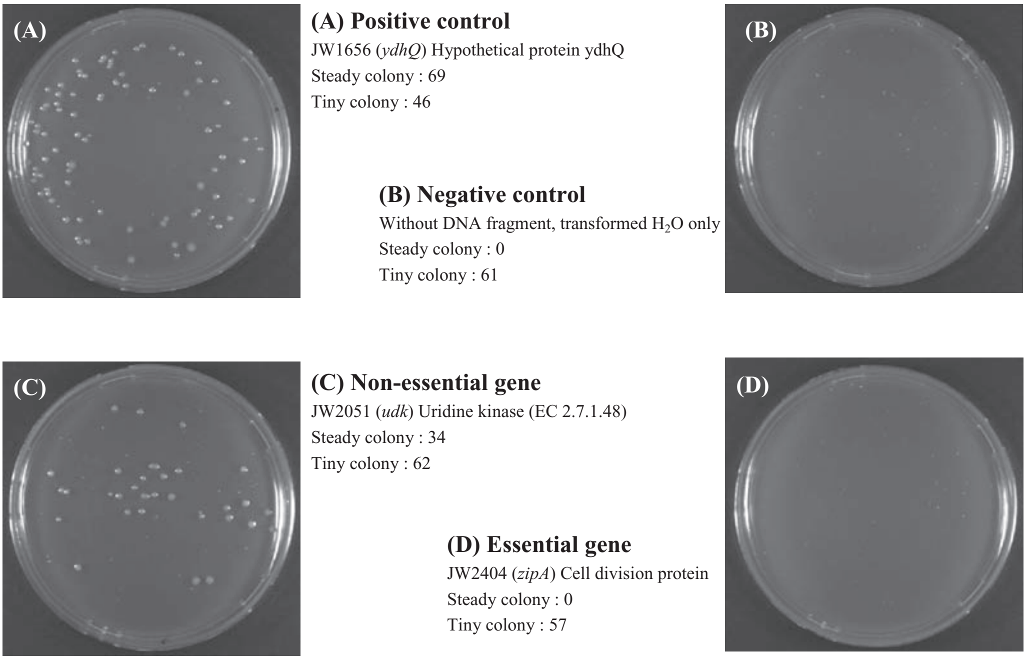


Fig. 5. E-Gel 96 High-Throughput Agarose Electrophoresis System (Invitrogen). The 96-well formatted electrophoresis corresponds with normal 96-well plates. **(A)** PCR reactions in 96-well format; **(B)** sample transfer from PCR plates to E-Gel 96 ready-made agarose gels by multichannel pipette or robotics; **(C)** multichamber electrophoresis system for high-throughput analysis within 12 min; **(D)** gel image capture and analysis using the editorial software provided with the system; **(E)** the edited gel image comparison with DNA marker lanes.



(A) Positive control
JW1656 (*ydhQ*) Hypothetical protein ydhQ
Steady colony : 69
Tiny colony : 46

(B) Negative control
Without DNA fragment, transformed H_2O only
Steady colony : 0
Tiny colony : 61

(C) Non-essential gene
JW2051 (*udk*) Uridine kinase (EC 2.7.1.48)
Steady colony : 34
Tiny colony : 62

(D) Essential gene
JW2404 (*zipA*) Cell division protein
Steady colony : 0
Tiny colony : 57

Fig. 6. Examples of transformants. Mutants were isolated in batches, in which each batch included a PCR product for disruption of ydhQ as a positive control and a no-PCR-product negative control. The latter usually gave only 10 to 100 tiny colonies.

Table 1
Genomic PCR Conditions

		Touchdown PCR conditions			Constant PCR conditions			Additional
PCR conditions	Hot start	Denature	Annealing	Extension	Denature	Annealing	Extension	extension
Normal		—			30 cycles			
	95.0	—	—	—	94.0	60.0	72.0	72.0
	2:00	—	—	—	0:30	0:30	5:00	2:00
Medium		8 cycles (–0.5°C/cycle)			25 cycles			
	95.0	94.0	70.0	72.0	94.0	66.0	72.0	—
	2:00	0:30	0:30	9:00	0:30	0:30	9:00	—
High		8 cycles (–0.5°C /cycle)			25 cycles			
	95.0	94.0	72.0	72.0	94.0	68.0	72.0	—
	2:00	0:30	0:30	9:00	0:30	0:30	9:00	—
Low		8 cycles (–0.5°C /cycle)			25 cycles			
	95.0	94.0	66.0	72.0	94.0	62.0	72.0	—
	2:00	0:30	0:30	9:00	0:30	0:30	9:00	—

Upper, temperature; lower, reaction time (min:s).

17. We designed several sets of PCR conditions for verification, including "normal" and various "touchdown PCR" conditions (**Table 1**). "High," "Medium," or "Low" touchdown PCR conditions were selected to match the Tm of gene specific primers (**Fig. 3**). They were also used if no amplification was achieved by "normal" genomic PCR or if multiple nonspecific bands were observed with particular gene-specific primers.

Acknowledgments

The authors thank Dr. B.L. Wanner for providing the bacterial strains and plasmids. This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology of Japan and a grant from CREST, JST (Japan Science and Technology), NEDO (New Energy and Industrial Technology Development Organization), Tsuruoka City and Yamagata prefecture in Japan.

References

1. Datsenko, K. A., and Wanner, B. L. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6640–6645.
2. Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391.
3. Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., Arnaud, M., et al. (2003) Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4678–4683.
4. Jacobs, M. A., Alwood, A., Thaipisuttikul, I., Spencer, D., Haugen, E., Ernst, S., et al. (2003) Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 14339–14344.
5. Mori, H., Isono, K., Horiuchi, T., and Miki, T. (2000) Functional genomics of *Escherichia coli* in Japan. *Res. Microbiol.* **151**, 121–128.
6. Kohara, Y., Akiyama, K., and Isono, K. (1987) The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50**, 495–508.
7. Sambrook, J., and Russell, D. W. (2001) *Molecular Cloning, a Laboratory Manual,* third ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
8. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., et al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knock-out mutants – the Keio collection. *Mol. Systems Biol.* **2**, 8.
9. Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., et al. (2006) Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Mol. Systems Biol.* **2**, 7.
10. Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K., Blattner, F. R., Chaudhuri, R. R., et al. (2006) *Escherichia coli* K-12: a cooperatively developed annotation snapshot–2005, *Nucleic Acids Res.* **34**, 1–9.
11. Kitagawa, M., Ara, T., Arifuzzaman, M., Ioka-Nakamichi, T., Inamoto, E., Toyonaga, H., and Mori, H. (2005) Complete set of ORF clones of *Escherichia coli* ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA Res.* **12**, 291–299.

12

The Applications of Systematic In-Frame, Single-Gene Knockout Mutant Collection of *Escherichia coli* K-12

Tomoya Baba, Hsuan-Cheng Huan, Kirill Datsenko, Barry L. Wanner, and Hirotada Mori

Key Words: *Escherichia coli* K-12; essential genes; Keio collection.

1. Introduction

The increasing genome sequence data of microorganisms has provided the basis for comprehensive understanding of organisms at the molecular level. Besides sequence data, a large number of experimental and computational resources are required for genome-scale analyses. *Escherichia coli* K-12 has been one of the best characterized organisms in molecular biology. Recently, the whole-genome sequences of two closely related *E. coli* K-12 strains, MG1655 *(1)* and W3110 *(2)*, were compared and confirmed by resequencing selected regions from both strains *(2)*. The availability of highly accurate *E. coli* K-12 genomes provided an impetus for the cooperative reannotation of both MG1655 and W3110 *(3)*. A set of precisely defined, single-gene knockout mutants of all nonessential genes in *E. coli* K-12 was constructed based on the recent accurate genome sequence data (*[4]* and **Chapter 11**). These mutants were designed to create in-frame (nonpolar) deletions upon elimination of the resistance cassette. These mutants have provided new key information on *E. coli* biology. First, the vast majority of the 3985 genes that were independently disrupted at least twice are probably nonessential, at least under the conditions of selection. Second, the 303 genes that we repeatedly failed to disrupt are candidates for *E. coli* essential genes. Lastly, phenotypic effects of all these mutations in the uniform genetic background of *E. coli* BW25113 were assessed by profiling mutants' growth yields on rich and minimal media *(4)*. These mutants should provide not only a basic resource for systematic functional genomics but also an experimental data source for systems biology applications. The mutants can serve as fundamental tools for a number of reverse genetics approaches, permitting analysis of the consequences of the complete loss of gene function, in contrast with forward genetics approaches in which mutant phenotypes are associated with a

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

corresponding gene or genes. Providing this resource to the research community will contribute to worldwide efforts directed toward a comprehensive understanding of the *E. coli* K-12 model cell. Because many *E. coli* gene products are well conserved in nature, these mutants are likely to be useful not only for studying *E. coli* and other bacteria but also for examining properties of genes from a wide range of living organisms.

2. Materials

1. Essential genes list by Gerdes et al. (*5*): http://www.integratedgenomics.com/online_material/gerdes/.
2. Essential genes from the PEC database and related documents (*[6]* and (**Chapter 26**): http://shigen.lab.nig.ac.jp/ecoli/pec/.
3. Potentially essential genes list by Kang et al. (*7*): http://jb.asm.org/cgi/content/full/186/15/4921/DC1.
4. COG database and related documents (*8*): http://www.ncbi.nlm.nih.gov/COG/new/.
5. Microbial Genome Database (*9*): http://mbgd.genome.ad.jp/.
6. G-language system and related documents (*10*): http://www.g-language.org/.

3. Methods

3.1. Evaluation of Gene Essentiality in E. coli *K-12*

One way to evaluate gene essentiality is to examine the efficiency of *E. coli* transformation with gene-specific linear knockout constructs generated according to Wanner's one-step polymerase chain reaction (PCR)-based gene inactivation protocol, as described in Ref. *4* and **Chapter 11**. Briefly, *E. coli* strain BW25113 carrying plasmids pKD20 or pKD46 (a standard strain for Wanner's one-step inactivation method; *see* **Chapter 11**) were propagated in the presence of L-arabinose to induce production of λ RED recombinase and subsequently transformed with PCR-generated mutagenic constructs specific for each targeted gene. The resultant transformants (eight for each gene target) were analyzed for the presence of kanamycin-resistance gene in the expected chromosomal location using several sets of *kan*-specific and locus-specific primers (**Fig. 1**; *see* also **Chapter 11**).

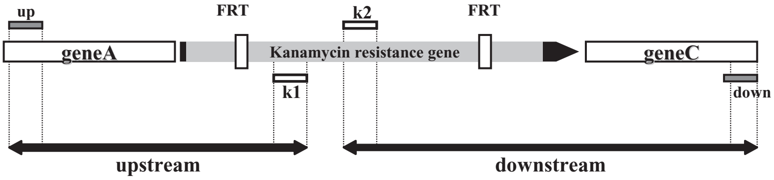


Fig. 1. Sample deletion mutant (Δ*geneB*) and the scheme of PCR-based verification of the postdeletion structure of the *kan* cassette and up- or downstream genes in the immediate vicinity of a targeted gene. Novel junctions created between the resistance cassette and neighboring upstream (*geneA*) and downstream (*geneC*) sequences are verified by PCR amplification using kanamycin-specific (k1 or k2) and locus-specific (up or down) primers.

Table 1
Gene Knockout Efficiency*

		Essentiality score[‡]		
Percentage correct[†]	ORFs	<–1	–1 to +1	>1
100	1946	1916	30	0
87.5	729	719	10	0
75	499	487	12	0
62.5	316	307	9	0
50	219	211	8	0
37.5	116	112	4	0
25	160	149	11	0
12.5	1	0	1	0
0	302	0	88	214
Total	4288	3901	173	214

*The original data are presented in the Supplementary Table 3 ***(4)***.

[†]Percentage of sample Km^R transformant (out of four or eight tested) shown by PCR analysis (**Section 3.1** and **Chapter 11**) to have the correct deletion structure.

[‡]The number of ORFs with different essentiality scores is given. Scores less than −1 identify a gene as nonessential and greater than +1 as essential with no contradictions with previous studies. Scores between −1 and +1 mean some inconsistency exists.

Out of all transformants tested, nearly 77% had the expected structure for the correct deletion. Out of the 4288 targeted open reading frames (ORFs), at least 50% of Km-resistant transformants were correct for 3490 constructs (**Table 1**). We tentatively asserted an ORF as essential if 7 or 8 of 8 analyzed clones failed to demonstrate the expected locus structure, as tested by PCR.

3.2. Functional Categories of Essential Genes

Clusters of orthologous groups (COGs) of proteins provide us not only with evolutionary information but also with functional information ***(8, 11)***. ORFs can be classified into COGs belonging to different functional categories (**Fig. 2**). The easiest way to identify the COGs category of a target ORF is to access the COGnitor page (http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html), paste an amino acids sequence into the query box, and compare it to the COGs sequences. Although this is easy to use, only a single sequence can be analyzed at a time, and a script needs to be developed in order to analyze multiple amino acids sequences. Required steps are

1. Prepare an amino acids sequence library in multiFASTA format.
2. Compare it against the full prokaryotic protein database *xyva* (available at ftp://ftp.ncbi.nih.gov/pub/COG/old/) using BLAST (blastp) analysis.
3. Use Dignitor, the program for identification of COGs in batch mode; requires the specific format shown in **Figure 3A** modified from blastp output files.
4. Run Dignitor to classify multiple proteins into COGs.

G-language is an environment for genome analysis developed by Arakawa et al. ***(10)***. COG analysis is also covered by G-language, and the script shown in **Figure 3B**

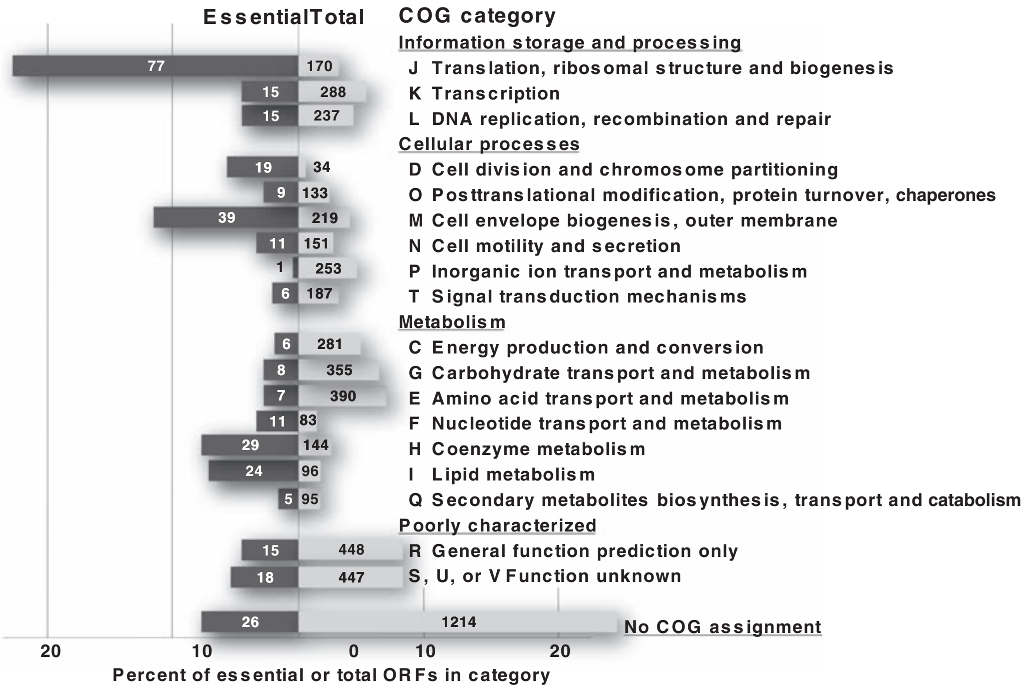


Fig. 2. COG classification of essential (white numbers on black bars) and of all (black numbers on gray bars) *E. coli* K-12 genes.

generates COG ID, functional category, product, number of protein hits, and BeTs (the protein in a target genome, which is most similar to a given protein from the query genome) score automatically.

Some ORFs could be classified into multiple COGs, especially multidomain proteins. Some COGs also belong to more than one functional class. Consequently, the 303 essential ORF candidates correspond with 315 COGs (including 26 with no COG

(A) Format required for dignitor

```
prot1 - HP1114     (746 1e-79)     10..399  3..398?
                   (230 4e-19)     474..555  473..554?
prot2 - sll1525    (158 3e-11)     3..184  6..183
```

(B) Sample script of G-language for COG classification

```
$gb = new G("multifasta protein library file name");
do{
        @result = cognitor($gb->{SEQ});
        printf" %s: %s\n" , $gb->{LOCUS}->{id}, $result[0];
}while($gb->next_locus());
```

Fig. 3. Sample script in G-language for classification of multiple proteins into COGs.

assignment) (**Fig. 2**). The fraction of essential genes varies widely with the COG classification. The greatest fractions are in translation, ribosomal structure, and biogenesis. The vast majority of essential genes are associated with cell division, lipid metabolism, translation, transcription, and cell envelope biogenesis. For example, our results indicate that *rpoE* and *rpoH*, encoding the RNA polymerase heat shock sigma factors E and H, respectively, are essential, in agreement with earlier studies ***(12, 13)***. Our data also showed that we were able to disrupt genes for five ribosomal proteins (S6, S20, L1, L11, and L33), which had been previously shown to be nonessential ***(14)***. Discrepancies for 11 others may have resulted from the use of different growth conditions, strain, or some other artificial problem, such as accumulation of suppressor mutations described above.

3.3. Comparison with Essential Genes Detected by Other Methods and Other Bacteria

Several systematic approaches for identification of essential genes have been performed. Genetic footprinting ***(5, 15)*** revealed 620 genes to be essential for robust aerobic growth of *E. coli* K-12. Yet, only 67% (205 genes) overlap with the predicted essential genes in this study. Striking differences can be attributed to the use of different mutagenesis strategies (transposon insertion vs. deletion), different growth conditions (broth vs. agar), or the approach for discriminating essential versus nonessential genes. Because genetic footprinting measures cell populations, a mutation causing mild reduction in growth rate can lead to the underrepresentation of a mutant in population and, hence, false classification of the corresponding gene as essential. In contrast, we sought deletion mutants as survivors without regard to growth rate. Comparisons of our results with those from genetic footprinting ***(5)***, the PEC database ***(6)***, and transposon mutagenesis ***(7)*** were performed and "essentiality scores" computed for all 303 essential gene candidates from our study ***(4)***. We also examined the conservation of the *E. coli* K-12 essential genes in genomes of other organisms deposed in the Microbial Genome Database (http://mbgd.genome.ad.jp/ ***[9]***). Comparison with three other *E. coli* genomes revealed that more than 90% (282) of the essential genes are universally present. About one-half (147) are conserved among 20 different Enterobacteriaceae genomes. One-third (85) are conserved among 74 Proteobacteria, and less than 15% (42) are conserved among 171 bacteria ***(4)***. *Bacillus subtilis* has a 4.2-MB genome and 271 essential genes, as determined by creation of the systematic gene knockout library ***(16)***. About one-half (150) of the orthologous genes are also essential in *E. coli*. Another 67 genes that are essential in *E. coli* are not essential in *B. subtilis,* and 86 *E. coli* essential genes have no *B. subtilis* ortholog. Details are reported elsewhere ***(4)***.

3.4. Growth Profiling of Rich and Minimal Media

All mutants were analyzed for growth yield by the 96-well optical photometer (SPECTRAmax PLUS, Molecular Devices, Eugene, OR) in both rich (LB) and minimal glucose MOPS ***(17)*** media after 22 and 48 h, respectively (**Fig. 4**). Growth data in **Figure 4** are summarized according to COG, and the complete information is available in Supplementary Table 3 in our manuscript ***(4)***. All mutants in Keio collection (spots on the plot) were grouped into seven groups (from I to VII) based on the average growth

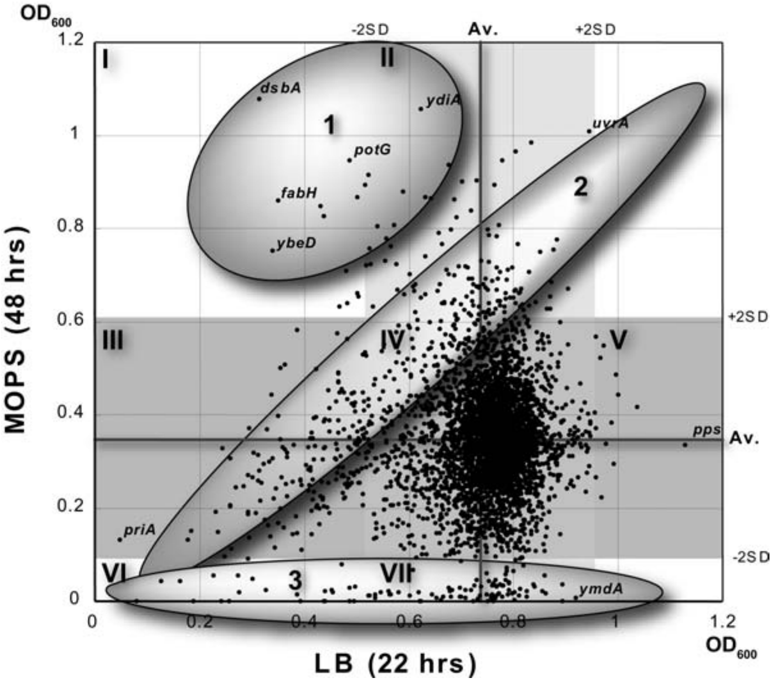


Fig. 4. Profiling gene contributions for growth. Mutants of all 3985 genes in the Keio collection were grown for 22 h in LB and 24 and 48 h in 0.4% glucose MOPS 2 mM P_i medium ***(17)***. Maximal cell density values obtained are plotted. Circled areas 1, 2, and 3 are discussed in the text (**Section 3.5**). Grayed areas show 2× standard deviation from the maximal cell density obtained for the wild-type strain. Groups labeled I to VII differ by more than 2× standard deviations.

yield and standard deviation on each axis. The vast majority appeared no different from the wild-type (group IV in **Fig. 4**). Mutants in circled **area 1** gave higher yield in minimal medium than in rich one; those in **area 2** gave similar yields in both medium; and those in **area 3** failed to grow in minimal medium. As expected, the majority of mutants in **area 3** have defects in biosynthesis of amino acids, purines, pyrimidines, or vitamins. Curiously, a subset of these auxotrophs showed modest growth after 48 h, suggesting that suppressors arose. A few mutants with deletions of genes with unknown function also grew well in rich but not in minimal medium, which may provide a handle on determination of their function. Some grew after 24 h in minimal medium but showed no growth after 48 h, suggesting possible cell *lysis* (*see* **Note 2**), the majority of mutant strains showed no striking growth defects.

3.5. Protein-Protein Interaction Network Profile of Essential Gene Products

As protein-protein interactions are central to most biological processes, uncovering large-scale properties of protein interaction networks potentially offers a deeper understanding of the system-level properties of living organisms ***(18)***. The topological parameters used to compare and characterize the protein interaction networks are as follows (*19*):

1. The degree *K* (also called average connectivity), which indicates how many links a given node has with the others (*19*).
2. The clustering coefficient *C*, a measure of the connectivity of a node. It corresponds with the fraction of the existing links compared with all possible links of a node (*20*).
3. The mean path length *L*, which indicates the average of the distance (the smallest number of links that we have to pass through to travel between two nodes) between all pairs of nodes (*18*).
4. The diameter *D*, which indicates the maximum internode distance (*21*).
5. The between-ness centrality *B*, which characterizes the degree of influence a protein has in “communicating” between protein pairs and is defined as the fraction of shortest paths going through a given node.

It has been shown that some protein interaction networks follow power-law distributions; that is, they consist of many interconnecting nodes, a few of which have uncharacteristically high degrees of connectivity (hubs). In addition, power-law distributions can be characterized as scale-free; that is, the possibility for a node to have a certain number of links does not depend on the total number of nodes within the network (i.e., the scale of the network). Scale-free networks provide stability to the cell because many non-hub genes can be disabled without greatly affecting the viability of the cell. Recently, Jeong and co-workers (*20*) focused on the relationship between hubs and essential genes and determined that hubs tend to be essential.

We constructed a comprehensive *E. coli* protein interaction network containing 11,511 unique interactions among 3047 proteins based on the large-scale experimental measurement (*22*). The topological characteristics between the essential gene products and the nonessential ones are compared as listed in **Table 2**. In a gross comparison, we found that essential gene products have significantly more links than the nonessential gene products, validating earlier findings in budding yeast (*20, 21*). Specifically, essential gene products have approximately twice as many links compared with nonessential gene products. We can also see from the power-law plots of the interactions of essential and nonessential gene products (**Fig. 5**) that the essential gene products have

Table 2
Topological Properties of the *E. coli* K-12 Protein-Protein Interaction Network

	Essential	Nonessential	*p* value
Average degree (*K*)	17.97	6.67	$<10^{-9}$
Average clustering coefficient (*C*)	0.058	0.064	0.081
Mean path length (*L*)	3.235	3.376	$<10^{-16}$
Diameter (*D*)	8	9	—
Average between-ness centrality (*B*)	0.00251	0.00063	$<10^{-4}$

Comparison of topological characteristics of essential gene products and nonessential gene products in the protein-protein interaction network of *E. coli* K-12. The *p* values are calculated using Wilcoxon rank sum test.

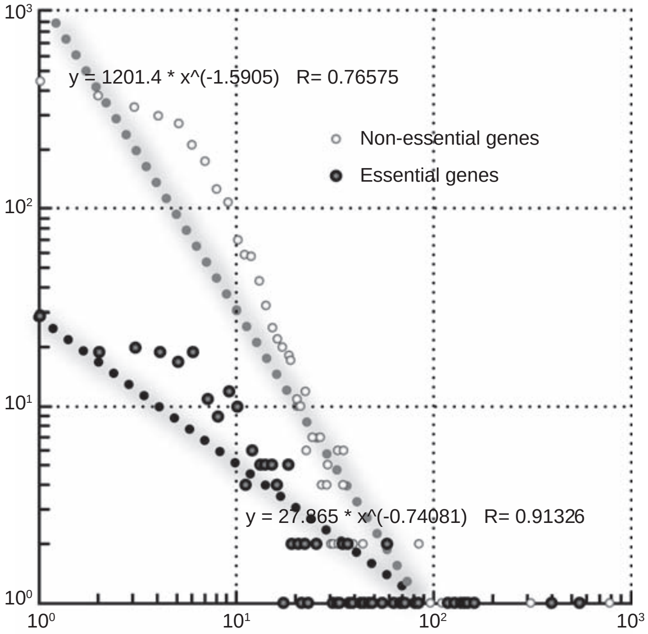


Fig. 5. Topological properties of the *E. coli* protein-protein interaction network: the comparison of the frequency of interacting partner proteins between essential and nonessential gene products. The *x*-axis represents the number of protein partners, and the *y*-axis represents the frequency of interactions.

a shallower slope, indicating that a proportionately larger fraction of them are hubs. Furthermore, within the interaction network, essential genes tend to be more closely connected to each other as determined from the mean path length and diameter. The between-ness centrality also implies that the essential gene products play a more important role in the network.

3.6. Use and Distribution of Knockout Mutants

Several complete sets of the Keio collection as well as thousands of individual mutants have already been distributed worldwide. Distribution is being handled via GenoBase (http://ecoli.naist.jp/) and National BioResource Project (http://shigen.lab.nig.ac.jp/ecoli/strain/top/top.jsp; *see* also **Chapter 26**) together with supporting data and other key resources, including the ASKA (a complete set of *E. coli* K-12 ORF archive) clone set (*22*). Several studies have already reported use of these mutants. For example, single-gene deletion mutants of the Keio collection were utilized for the study of uncharacterized gene function (*23*) and the analysis of metabolism (*24–29*). The use of subsets of Keio collection mutants has substantiated the value of systematic approaches for the understanding of cellular systems (*30–32*). Construction of deletion mutants of essential genes in the presence of the corresponding wild-type alleles provided on a plasmid in trans are now underway. They will become an open resource for the community as well.

Notes

1. Limitations of identification of essential genes by transformation efficiency alone: There are several limitations to identifying essential genes by our method. For example, in case of *secM*, just one candidate clone out of eight tested had the expected structure, with this outcome reproducible in several independent experiments. In this exceptional case, *secM* harbors a translational arrest sequence within its C-terminus that is required for expression of the downstream *secA*, encoding an essential preprotein translocase SecA subunit (*33, 34*). Thus, it is reasonable to suggest that the sole *secM* mutant arose because it acquired a suppressor allowing *secA* expression. The ability to select directly for knockout mutants may have led to other mutants with suppressors. For example, the same mutagenesis strategy has been used elsewhere to create a deletion of *mreB* (*35*), an essential gene, in which case, the mutant was later shown to carry a suppressor (*36*). Yet, we repeatedly failed to recover a Δ*mreB* mutant, even when using the primers and host strain identical to those in (*35*). We have also confirmed the absence of the *mreB* coding sequences in their Δ*mreB* mutant isolated in this study, thus ruling out the possibility of a duplicate *mreB* sequence (data not shown). Clearly, *secM* and *mreB* are examples of "quasi-essential" genes, when suppressors allow viability of mutants with the respective deletions. By definition, deletion of truly essential genes cannot be mutationally suppressed. In addition to suppressors, a functional redundancy or duplication can obscure gene essentiality. It is difficult to assess functional redundancy without further experimentation. However, gene duplications can explain why we recovered mutants with deletions of some genes, such as *ileS* and *glyS*, encoding isoleucyl-tRNA and glycine (b subunit) tRNA synthetases, which are known to be essential. In these cases, the mutants might carry intact copies of the respective deleted gene elsewhere (R. D'Ari and K. Nakahihashi, personal communication), presumably resulting from gene duplications. Nevertheless, because the vast majority of mutants were recovered at a high frequency (**Table 1**), neither suppressors nor duplications seem to be of major concern. Genetic duplications resulting from gene amplification have been well documented in bacteria; however, the frequency is low—under ordinary conditions about 1 in 400 genes on average is duplicated in a culture (*37*). If we assume similar values, then no more than about 10 of our mutants are likely to have a gene duplication altering the interpretation of results. Even though about 1.5% of the yeast mutants were eliminated due to duplications (*38*), most studies on gene essentiality fail to consider this issue. Genome sequence difference is another limitation of this method for identifying the essential genes. We have reported genome sequence conservation on the nucleotide level between two closely related strains of *E. coli* K-12 (*2*). At the same time, we also reported differences between them in the IS or phage-related sequence distribution. Because genome sequence of the BW25113 strain has not been determined, some genes might fail to be deleted due to the differences in the target sites, such as IS insertion. This can be solved by sequencing of the entire BW25113 genome, but this is not practical. Alternatively, confirmation of the genome structure of the border regions for the 303 candidate essential genes might solve this problem.
2. Growth profiling on rich and minimal media: Some knockout mutants showed no growth after 48 h even though they grew after 24 h, suggesting lysis such as *ddlB* (D-alanine:D-alanine ligase with OD_{600} at 24 h of 0.270; and OD_{600} at 48 h of 0.005), *csgC* (predicted curli production protein, 0.224 to 0.006), *rsxC* (predicted 4Fe-4S ferredoxin-type protein, 0.219 to 0.061), and others, such as *ymdA* (0.326 to 0.006). Many grew poorly in both rich and minimal media, for example, *priA* (primosome factor), *atp* (ATP synthase components), and *cyaA* (adenylate cyclase). Interestingly, some deletion strains showed better growth in minimal media than in rich media, such as *dsbA* (periplasmic protein disulfide isomerase I), *potG* (putrescine transporter subunit), fabH (3-oxoacyl-[acyl-carrier-protein] synthase III), and so forth.

Acknowledgments

We thank Miki Naba, Daisuke Kido, Chy Narith, Toru Kodama, Koji Komatsu, and Kazuyuki Shimizu from the Kyushu Institute of Technology for help in measuring the growth of the glycolysis gene deletion mutants. This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology of Japan; and a grant from CREST, JST (Japan Science and Technology), in part from NEDO (New Energy and Industrial Technology Development Organization) and from Tsuruoka city and Yamagata prefecture.

References

1. Blattner, F. R., Plunkett, G. 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., et al. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474.
2. Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., et al. (2006) Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Mol. Syst. Biol.*
3. Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K., Blattner, F. R., Chaudhuri, R. et al. (2006) *Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005. *Nucleic Acids Res.* **34**, 1–9.
4. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., et al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knock-out mutants—the Keio collection. *Mol. Syst. Biol.*
5. Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balazsi, G., Ravasz, E., Daugherty, M. D., et al. (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684.
6. Hashimoto, M., Ichimura, T., Mizoguchi, H., Tanaka, K., Fujimitsu, K., Keyamura, K., et al. (2005) Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. *Mol. Microbiol.* **55**, 137–149.
7. Kang, Y., Durfee, T., Glasner, J. D., Qiu, Y., Frisch, D., Winterberg, K. M. and Blattner, F. R. (2004) Systematic mutagenesis of the *Escherichia coli* genome. *J. Bacteriol.* **186**, 4921–4930.
8. Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997) A genomic perspective on protein families. *Science* **278**, 631–637.
9. Uchiyama, I. (2003) MBGD: microbial genome database for comparative analysis. *Nucleic Acids Res.* **31**, 58–62.
10. Arakawa, K., Mori, K., Ikeda, K., Matsuzaki, T., Kobayashi, Y., and Tomita, M. (2003) G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining. *Bioinformatics* **19**, 305–306.
11. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**, 22–28.
12. Hiratsu, K., Amemura, M., Nashimoto, H., Shinagawa, H., and Makino, K. (1995) The rpoE gene of *Escherichia coli*, which encodes sigma E, is essential for bacterial growth at high temperature. *J. Bacteriol.* **177**, 2918–2922.
13. Zhou, Y. N., Kusukawa, N., Erickson, J. W., Gross, C. A., and Yura, T. (1988) Isolation and characterization of *Escherichia coli* mutants that lack the heat shock sigma factor sigma 32. *J. Bacteriol.* **170**, 3640–3649.

14. Dabbs, E. R. (1991) Mutants lacking individual ribosomal proteins as a tool to investigate ribosomal properties. *Biochimie* **73**, 639–645.
15. Tong, X., Campbell, J. W., Balazsi, G., Kay, K. A., Wanner, B. L., Gerdes, S. Y., and Oltvai, Z. N. (2004) Genome-scale identification of conditionally essential genes in *E. coli* by DNA microarrays. *Biochem. Biophys. Res. Commun.* **322**, 347–354.
16. Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., Arnaud, M., et al. (2003) Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4678–4683.
17. Neidhardt, F. C., Bloch, P. L. and Smith, D. F. (1974) Culture medium for enterobacteria. *J. Bacteriol.,* **119,** 736–747.
18. Yook, S. H., Oltvai, Z. N., and Barabasi, A. L. (2004) Functional and topological characterization of protein interaction networks. *Proteomics* **4**, 928–942.
19. Barabasi, A. L., and Oltvai, Z. N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113.
20. Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001) Lethality and centrality in protein networks. *Nature* **411**, 41–42.
21. Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X., and Gerstein, M. (2004) Genomic analysis of essentiality within protein networks. *Trends Genet.* **20**, 227–231.
22. Arifuzzaman, M., Maeda, M., Itoh, A., Nishikata, K., Takita, C., Saito, R., et al. (2006) Large-scale identification of protein–protein interaction of *Escherichia coli* K-12. *Genome Res*. **16**, 686–691.
23. Melnick, J., Lis, E., Park, J. H., Kinsland, C., Mori, H., Baba, T., et al. (2004) Identification of the two missing bacterial genes involved in thiamine salvage: thiamine pyrophosphokinase and thiamine kinase. *J. Bacteriol.* **186**, 3660–3662.
24. Yang, C., Hua, Q., Baba, T., Mori, H., and Shimizu, K. (2003) Analysis of *Escherichia coli* anaplerotic metabolism and its regulation mechanisms from the metabolic responses to altered dilution rates and phosphoenolpyruvate carboxykinase knockout. *Biotechnol. Bioeng.* **84**, 129–144.
25. Hua, Q., Yang, C., Baba, T., Mori, H., and Shimizu, K. (2003) Responses of the central metabolism in *Escherichia coli* to phosphoglucose isomerase and glucose-6-phosphate dehydrogenase knockouts. *J. Bacteriol.* **185**, 7053–7067.
26. Hua, Q., Yang, C., Oshima, T., Mori, H., and Shimizu, K. (2004) Analysis of gene expression in *Escherichia coli* in response to changes of growth-limiting nutrient in chemostat cultures. *Appl. Environ. Microbiol.* **70**, 2354–2366.
27. Jiao, Z., Baba, T., Mori, H., and Shimizu, K. (2003) Analysis of metabolic and physiological responses to gnd knockout in *Escherichia coli* by using C-13 tracer experiment and enzyme activity measurement. *FEMS Microbiol. Lett.* **220**, 295–301.
28. Zhao, J., Baba, T., Mori, H., and Shimizu, K. (2004) Effect of zwf gene knockout on the metabolism of *Escherichia coli* grown on glucose or acetate. *Metab. Eng.* **6**, 164–174.
29. Zhao, J., Baba, T., Mori, H., and Shimizu, K. (2004) Global metabolic response of *Escherichia coli* to gnd or zwf gene-knockout, based on 13C-labeling experiments and the measurement of enzyme activities. *Appl. Microbiol. Biotechnol.* **64**, 91–98.
30. Perrenoud, A., and Sauer, U. (2005) Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in *Escherichia coli. J. Bacteriol.* **187**, 3171–3179.
31. Tenorio, E., Saeki, T., Fujita, K., Kitakawa, M., Baba, T., Mori, H., and Isono, K. (2003) Systematic characterization of *Escherichia coli* genes/ORFs affecting biofilm formation. *FEMS Microbiol. Lett.* **225**, 107–114.

32. Itoh, A., Ohashi, Y., Soga, T., Mori, H., Nishioka, T., and Tomita, M. (2004) Application of capillary electrophoresis-mass spectrometry to synthetic *in vitro* glycolysis studies. *Electrophoresis* **25**, 1996–2002.
33. Nakatogawa, H., Murakami, A., Mori, H., and Ito, K. (2005) *SecM* facilitates translocase function of *SecA* by localizing its biosynthesis. *Genes Dev.* **19**, 436–444.
34. Murakami, A., Nakatogawa, H., and Ito, K. (2004) Translation arrest of *SecM* is essential for the basal and regulated expression of *SecA*. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 12330–12335.
35. Kruse, T., Moller-Jensen, J., Lobner-Olesen, A., and Gerdes, K. (2003) Dysfunctional *MreB* inhibits chromosome segregation in *Escherichia coli*. *EMBO J.* **22**, 5283–5292.
36. Kruse, T., Bork-Jensen, J., and Gerdes, K. (2005) The morphogenetic *MreBCD* proteins of *Escherichia coli* form an essential membrane-bound complex. *Mol. Microbiol.* **55**, 78–89.
37. Anderson, R. P., and Roth, J. R. (1978) Tandem genetic duplications in *Salmonella typhimurium*: amplification of the histidine operon. *J. Mol. Biol.* **126**, 53–71.
38. Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391.

13

A Novel, Simple, High-Throughput Method for Isolation of Genome-Wide Transposon Insertion Mutants of *Escherichia coli* K-12

Takeyoshi Miki, Yoshihiro Yamamoto, and Hideo Matsuda

Summary

We developed a novel, simple, high-throughput method for isolation of genome-wide transposon insertion mutants of *Escherichia coli* K-12. The basic idea of the method is to randomly disrupt the genes on the DNA fragments cloned on the Kohara library by inserting a mini-transposon first, and then transfer the disrupted genes from the λ vector to the *E. coli* chromosome by homologous recombination. Using this method, we constructed a set of 8402 Km^r *cis*-diploid mutants harboring a mini-Tn*10* insertion mutation and the corresponding wild-type gene on a chromosome, as well as a set of 6954 haploid mutants derived from the *cis*-diploid mutants. The major advantage of the strategy used is that the indispensable genes or sites for growth can be identified. Preliminary results suggest that 415 open reading frames are indispensable for growth in *E. coli* cells. A total of 6404 haploid mutants were deposited to Genetic Strains Research Center, National Institute of Genetics, Japan (**Chapter 26**) and are available for public distribution upon request (http://shigen.lab.nig.ac.jp/ecoli/strain/nbrp/resource.jsp).

Key Words: *Escherichia coli*; indispensable gene; insertion mutant; transposon.

1. Introduction

As a model organism, *Escherichia coli* has been playing significant role in the establishment of a number of basic concepts in molecular biology, and the enormous amount of data accumulated to date has contributed to the understanding of a variety of cellular processes. Nevertheless, the function of about half of the 4300 open reading frames identified by DNA sequencing of the whole genome were unknown. As a first step to execute systematic function analysis of *E. coli* genes, we developed a simple, high-throughput method for isolation of genome-wide transposon insertion mutants of *E. coli* K-12. The purpose of this chapter is to describe in some detail both the methods developed here and the most recent collections of mini-Tn*10* insertion mutants of *E. coli* K-12.

The basic idea of the method is to randomly disrupt the genes located on DNA fragments cloned in the Kohara library (*1*) by inserting a mini-transposon; and then to

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

introduce disrupted genes into the *E. coli* chromosome by homologous recombination. Because Kohara clones (a collection of ordered lambda clones carrying *E. coli* DNA segments *[1]*) cover practically the whole *E. coli* genome, this method should allow genome-wide isolation of insertion mutants.

As shown in **Figure 1**, the method consists of four steps. The first step is to generate a random mutant sublibrary for each Kohara clone by propagating it in a host strain carrying the mini-Tn*10*(*lacZ*α*-kan*) donor plasmid. The second step is to lysogenize mutated phage sublibraries via recombination between the cloned insert and the homologous region on the *E. coli* chromosome (i.e., to construct partial diploids harboring both wild-type and disrupted alleles on a chromosome). The only requirement for this step is to supply the *cI* repressor, as Kohara clones have the *cI* gene deleted. The lack of the *int* gene in the Kohara clones facilitates homologous recombination by suppressing the integration at lambda attachment sites on the chromosome. The third step is to select nonlysogenic insertion mutants produced spontaneously by recombination between the duplicated regions (i.e., to construct haploid disruption mutants from the partial *cis*-diploids). The last step is to map the introduced transposon insertions by sequencing across transposon-chromosome boundaries and to identify disrupted genes.

The major advantage of this strategy is that the genes or sites indispensable for *E. coli* growth and survival can be identified. The *cis*-diploid cells harboring both a wild-type allele and a disrupted one are expected to be viable even if the insertion occurred within a gene (or site) indispensable for growth. In contrast, in a haploid strain insertion in a gene indispensable for growth will be lethal. Hence, by testing whether haploid disruption mutants arise from the *cis*-diploid cells, genes essential for growth can be identified.

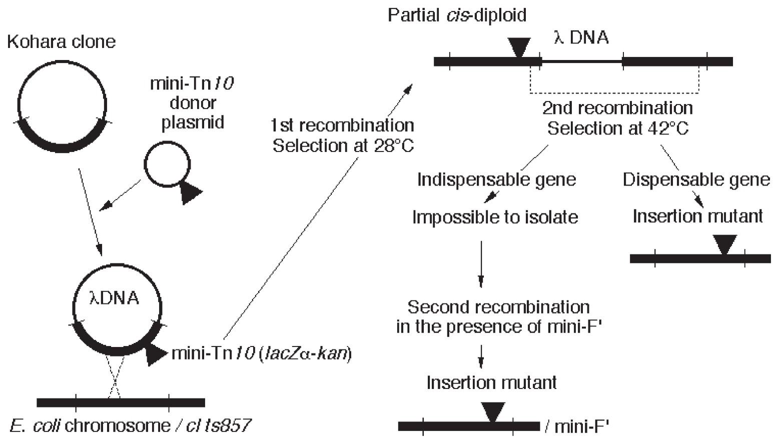


Fig. 1. Schematic diagram of the systematic disruption of *E. coli* chromosome using mini-Tn*10*(*lacZ*α*-kan*) transposon mutagenesis of Kohara clones. See text for detail.

2. Materials

2.1. Reagents and Labware

1. Kanamycin: Kanamycin monosulfate.
2. Ampicillin: D[–]-α-Aminobenzylpenicillin.
3. Chloramphenicol.
4. IPTG: Isopropyl-β-D(–)-thiogalactopyranoside.
5. TaKaRa LA PCR Kit (Takara Bio Inc. Otsu, Japan).
6. Automated thermal cycler GeneAmp PCR system 9700 (Applied Biosystems, Foster City, CA), used under 9600 mode.
7. Polymerase chain reaction (PCR) reaction plates: MicroAmp Optical 96 Well Reaction Plates and MicroAmp Caps (Applied Biosystems, Foster City, CA; catalogue no. N801-0560 and N801-0535).
8. PCR product Pre-Sequencing Kit: Exonuclease I (10 U/μL) and shrimp alkaline phosphatase (2 U/μL) (USB Corporation, Cleveland, OH).
9. BigDye Terminator Cycle Sequencing Kit: Version 1.1 (Applied Biosystems).

2.2. Buffers

1. TMG buffer: Composition per liter of distilled water: 1.2 g Tris base, 2.5 g $MgSO_4$ $7H_2O$, 0.1 g gelatin. Adjust to pH 7.4 with HCl, heat to dissolve gelatin, and autoclave.
2. Elution buffer: Composition per liter of water: 3.0 g Tris base, 0.25 g $MgSO_4$ $7H_2O$. Adjust to pH 8.0 with HCl.

2.3. Media

1. Tryptone broth: Per liter distilled water: 10 g Bacto Tryptone, 5 g NaCl.
2. TBMM: Tryptone B1 broth with maltose and magnesium. Per liter distilled water: 10 g Bacto Tryptone, 5 g NaCl; autoclave, add filtered maltose to final concentration of 0.2% (w/v) and $MgSO_4$ from sterile stock to final concentration of 10 mM. Add filter-sterilized thiamin to final concentration of 1 μg/mL.
3. TB1 agar plates: Make up Tryptone broth; add 11 g/L Bacto agar before autoclaving; add $MgSO_4$ to final concentration of 10 mM and thiamin to final concentration of 1 μg/mL; pour into Petri plates when agar is partially cooled. For titration of phages, use the plates while relatively fresh (within 1 week). For replica plating, dry the plates overnight at about 40°C before use.
4. Top agar: Tryptone broth with 7 g/L Bacto agar.
5. LB (Luria broth): Per liter distilled water: 10 g Bacto Tryptone, 5 g yeast extract, 5 g NaCl. Adjust to pH 7.5.
6. LBCit broth: Per liter distilled water: 10 g Bacto Tryptone, 5 g yeast extract, 5 g NaCl, 5 g Na_3 citrate dihydrate. Adjust to pH 7.5 (**Note 1**).
7. LBCit agar plates: Make up LBCit broth; add 15 g/L agar before autoclaving. When agar is partially cooled, pour into Petri dishes. Dry plates overnight at about 40°C before use.
8. TYCitSucrose agar plates: Per liter distilled water: 10 g Bacto Tryptone, 5 g yeast extract, 5 g Na_3 citrate dihydrate, 50 g sucrose. Adjust to pH 7.5. Add 15 g/L agar before autoclaving. Pour into Petri plates, when partially cooled. Dry plates overnight at about 40°C before use.
9. Antibiotics: Use at the following concentrations: kanamycin to final concentration of 25 μg/mL, ampicillin 25 μg/mL, chloramphenicol 12.5 μg/mL.
10. 80% glycerol: 65% glycerol (v/v), 0.1 M $MgSO_4$, 0.025 M TrisCl, pH 8.

2.4. Cultureware

1. Rectangular Petri dishes: omnitrays with lid (Nunc A/S, Roskilde, Denmark).
2. Microwell plate: U96 microwell plates (Nunc).
3. Replication system: Nunc Replication System (Nunc, catalogue no. 250520 and 250555).
4. Rectangular replication block.

2.5. Bacterial Strains, Plasmids, and Bacteriophages

1. Bacterial strains:
   KP7600 (T. Miki, unpublished): F$^-$ *lacI^Q lacZΔM15 galK2 galT22* λ$^-$ IN*(rrnD-rrnE)1*. A derivative of W3110(A) *(2)*.
   LE392: F$^-$ *e14$^-$(Mcr$^-$) hsdR514(r$_K^-$m$_K^+$) glnV44 supF58 lacY or Δ(lacZY)6 galK2 galT22 melB1 trpR55* ***(3)***.
2. Donor strain: KP7600 harboring pKP2371 (T. Miki, unpublished) and pKP2373 (T. Miki, unpublished). pKP2371 is a mini-Tn*10*(*lacZ*α*-kan*) donor plasmid with Cm^r marker (**Note 2**). pKP2373 is a pHG329 derivative harboring the *lacI*^{Q} gene.
3. Recipient strain: KP7600 harboring pKP2374 (T. Miki, unpublished). pKP2374 is an Ap^r mini-R plasmid harboring the *cIts857* gene of λ phage and the *sacB* gene of *B. subtilis* (**Note 3**).
4. The Kohara bacteriophage λ miniset collection ***(1)***, based on the EMBL4 λ vector ***(4)***.

2.6. Primers

1. PCR primers (**Note 4**):
   LAM1: 5′-ACAGTCGGTGGTCCGGCAGTACAATGGATTACC-3′.
   LAM2: 5′-GCAACCTGCAACGTATTGAGCGCAAGAATCAGC-3′.
2. Sequencing primer (**Note 5**):
   TP3: 5′-CGACGTTGTAAAACGACGGCCAGT-3′.

3. Methods

3.1. Isolation of Genome-Wide Transposon Insertion Mutants

3.1.1. Generating mini-Tn10(lacZa-kan) Insertion Sublibraries for Each Kohara Clone

1. Grow the donor strain to saturation in TBMM at 37°C.
2. Collect the cells by centrifugation and suspend the pellet in 10 mM $MgSO_4$ solution to OD_{600} of 1.0.
3. Add 2×10^6 target phage particles (originating from a single Kohara clone) to 0.1 mL of donor strain and allow to adsorb for 20 min at 37°C.
4. Dilute with 4.0 mL LB supplemented with 10 mM $MgSO_4$ and incubate at 37°C with vigorous shaking.
5. In 100 min, add 1/100 volume of 0.1 M IPTG to induce the transposase expression.
6. After lysis has occurred, add a few drops of chloroform, incubate for 15 min at 37°C, and pellet debris by centrifugation. Save supernatant, titer, aliquot, and store at 4°C.

3.1.2. Lysogenizing Sublibraries

1. Grow the recipient strain to saturation in LB supplemented with 0.2% maltose and 10 mM $MgSO_4$ at 28°C.

2. Collect the cells by centrifugation and suspend the pellet in 1/2 volume of 10 mM $MgSO_4$ solution.
3. Mix 1.6×10^{10} mutagenized phage particles (generated as described in **Section 3.1.1**) with 1.0 mL of the recipient strain and incubate 30 min at 28°C to allow adsorption.
4. Dilute the cell/bacteriophage mix with 10.0 mL LB supplemented with 25 μg/mL ampicillin and incubate for 2 h at 28°C for phenotypic expression.
5. Centrifuge the culture to collect the cells and suspend the pellet in 1/2 to 1/3 volume of TMG buffer.
6. Spread the cell suspension onto LBCit plates supplemented with kanamycin and ampicillin. Incubate for 36 h at 28°C. About 300 to 500 colonies per 10^{10} PFU will be obtained (**Note 6**).
7. Using toothpicks, transfer 300 to 400 Km^r colonies onto LBCit plates supplemented with kanamycin and ampicillin and incubate for 18 h at 28°C.
8. Prepare TB1 plates overlaid with LE392 in advance. Grow LE392 to saturation in TBMM at 37°C, pellet and resuspend the cells in 10 mM $MgSO_4$ solution to bring OD_{600} to 1.0. Add 0.1 mL LE392 to 3.0 mL melted top agar (48°C), mix gently, and immediately pour onto a TB1 plate dried overnight at about 40°C in advance. Leave plates in a refrigerator for 2 to 4 h prior to use to harden the top agar by allowing bottom agar to absorb the water contained in the top agar.
9. Replica-plate the Km^r transductant plates onto:
   (a) LBCit plates supplemented with kanamycin and ampicillin (to preserve Km^r transductants).
   (b) LBCit plates supplemented with chloramphenicol (to eliminate clones with entire donor plasmid integrated).
   (c) TB1 plates overlaid with LE392 (to eliminate nonlysogenic haploid transductants). Incubate all LBCit plates at 28°C and TB1 plates overlaid with LE392 at 39°C (**Note 7**).
10. Select those colonies that formed lysis zones on TB1 plates but did not grow on chloramphenicol plates. Pick them onto rectangular LBCit plates supplemented with kanamycin and ampicillin with toothpicks. Incubate for 18 h at 28°C. Ninety-six or 192 Km^r colonies per each Kohara clone would be a good number to manage at a time.
11. Inoculate the Km^r colonies into 80 μL LBCit broth supplemented with kanamycin and ampicillin dispensed in microwell plates using the Nunc replication system. Incubate at 28°C overnight with mild shaking using a microwell plate shaker.
12. Add 80 μL of 80% glycerol, mix well, and store at −80°C.

3.1.3 Sequencing Across Transposon-Chromosome Junctions

1. Replica-plate the Km^R/Cm^S transductants, described in **Section 3.1.2, step 10**, onto rectangular TB1 plates overlaid with LE392 (prepared as described in **Section 3.1.2**, **step 8**) using a rectangular replication block. Incubate overnight at 39°C (**Note 7**).
2. Pick each lysis zone produced with a Pasteur pipette and wash out the plug in 100 μL elution buffer dispensed in microwell plates. Leave the plates for 1 to 2 h at room temperature to elute bacteriophages for PCR amplification. If necessary, they may be stored at −80°C.
3. Amplify *E. coli* DNA in each λ clone by polymerase chain reaction using a TaKaRa LA PCR Kit with PCR primers LAM1 and LAM2 (**Note 8**).

Reaction mixture:

Water	7.32 μL
10× Buffer, Mg free	2.0 μL
25 mM $MgCl_2$	1.88 μL
dNTPs, 2.5 mM each	3.2 μL
Primer LAM1, 10 μM	0.2 μL
Primer LAM2, 10 μM	0.2 μL
Phage suspension	5.0 μL
LA Taq polymerase (5 U/μL)	0.2 μL
Total	20.0 μL

Cycling condition:

Hot start	1 min at	95°C
Denature	1 min at	95°C
Anneal and extend	15 min at	68°C
30 cycles		
Time delay	10 min at	72°C
Soak	Overnight at	4°C

4. Analyze PCR products by agarose gel electrophoresis using 2 μL of each reaction mixture.
5. Digest unconsumed deoxynucleotide triphosphates (dNTPs) and primers by adding 1 μL shrimp alkaline phosphatase and 1 μL exonuclease I to 18 μL PCR mixtures and incubating at 37°C for 60 min, followed by incubation at 80°C for 20 min to inactivate the enzymes.
6. Determine the nucleotide sequences by the dideoxy chain-termination method of Sanger et al. *(5)*, using sequencing primer TP3. Use 4 μL PCR amplified DNA per sequencing reaction (in 10 μL total volume).
7. Determine the sites and orientations of transposon insertions in *E. coli* chromosome by comparing sequencing reads with the sequence of a parental Kohara clone.
8. Identify open reading frames (ORFs) where insertions have occurred.

3.1.4. Preparing cis*-Diploid Mutants Harboring a mini-Tn*10 *Insertion*

1. Select mutants harboring mutations in target ORFs.
2. Scrape the surface of glycerol stocks prepared in **Section 3.1.2**, **step 12**, and streak-purify the mutants to single colonies on LBCit plates supplemented with kanamycin and ampicillin at 28°C.
3. Prepare glycerol stocks of the purified mutants and store at −80°C.
4. Prepare phage lysate from each mutant, amplify bacterial DNA cloned in each λ clone, and determine the sequence of transposon-chromosome junctions, as described in **Section 3.1.3**, for confirmation.

*3.1.5. Preparing Haploid Mutants Harboring a mini-Tn*10 *Insertion*

1. Streak out the *cis*-diploid mutants obtained in **Section 3.1.4** on LBCit plates supplemented with kanamycin and ampicillin and incubate at 42.5°C overnight (**Note 7**).
2. Grow a single colony isolated from each *cis*-diploid mutant on LBCit plates supplemented with kanamycin (without ampicillin) at 37°C to allow spontaneous segregational loss of plasmid pKP2374 (**Note 9**).

Table 1
Plan for Plating *cis*-Diploid Mutants

Experiment no.	Dilution	No. plates to be used	Incubation
1	10^{-6}	2	28°C for 36 h
2	10^{-5}	2	42.5°C for 20 h
3	10^{-4}	2	42.5°C for 20 h
4	10^{-3}	2	42.5°C for 20 h
5	10^{-2}	2	42.5°C for 20 h

3. Streak out the culture on TYCit sucrose plates and incubate at 28°C to select for plasmid-free haploid insertion mutants (**Note 10**).
4. Purify the colonies obtained on LBCit plates at 37°C.
5. Prepare glycerol stocks of the purified mutants and store at –80°C.

3.2. Making Gene Essentiality Assertions (Notes 11 and 12)

1. Grow to saturation the *cis*-diploid mutants to be tested in LBCit medium supplemented with kanamycin and ampicillin at 28°C.
2. Prepare 10-fold serial dilutions of the culture in TMG buffer, spread 0.1 mL of each dilution on LBCit plates supplemented with kanamycin and ampicillin, and incubate at 28°C or 42.5°C, as indicated in **Table 1**.
3. Count the number of colonies and calculate the ratio of the colonies formed at 42.5°C over the ones formed at 28°C.
4. Assert the genes as dispensable or indispensable based on the frequencies of formation of Km^R haploid mutants, i.e., nonlysogenic derivatives of "*cis*-diploid" mutants (**Note 13**).

3.3. Whole Genome–Ordered Nonredundant Collection of *E. coli* mini-Tn*10* *cis*-Diploid Insertion Mutants Available for Public Distribution

We have constructed a genome-wide random-insertion library of an *E. coli* strain KP7600, a W3110 derivative, utilizing the strategy described here. Sublibraries of mini-Tn*10*(*lacZα-kan*) insertion mutants were constructed for each of 462 Kohara clones. A total of 135,000 independent Km^R lysogens (i.e., partial *cis*-diploid strains harboring a mini-Tn*10* insertion mutation and the corresponding wild-type gene on a chromosome) were constructed. By sequencing across transposon-chromosome junctions, 58,500 different insertions were mapped onto the *E. coli* chromosome. To construct a nonredundant library of *E. coli cis*-diploid transposon insertion mutants, two clones per nearly every *E. coli* ORF were selected, each harboring a mini-Tn*10*(*lacZα-kan*) insertion located near the 5′ end in one of the two opposite orientations. The selected clones were streak-purified to single colonies and resequenced for confirmation.

A total of 8402 mutants representing about 90% of the predicted *E. coli* ORFs were retained as "a *cis*-diploid collection" of mini-Tn*10* insertion mutants. Next, 6954 haploid derivatives were constructed from the "*cis*-diploid collection" mutants, tested, and retained as "a haploid collection." Both collections are available for public distribution upon request from Genetic Strains Research Center, National Institute of Genetics, Japan (http://shigen.lab.nig.ac.jp/ecoli/strain/nbrp/resource.jsp; *see* **Chapter 26**).

By testing whether haploid disruption mutants can be isolated from each *cis*-diploid mutant, indispensable *E. coli* genes were identified. Our preliminary results suggest that 370 ORFs are indispensable for *E. coli* growth on LBCit plates (**Section 2.3**) at 42.5°C.

Taking into account the 45 known essential genes, for which our attempts at isolating insertion mutants have been unsuccessful, the total of 415 genes were identified as potentially indispensable for normal growth of *E. coli* cells.

Notes

1. Sodium citrate prevents infection of nonlysogenic haploid derivatives with λ phages produced by lysogenic *cis*-diploid cells.
2. Plasmid pKP2371 is a pNK2887 derivative *(6)*. The *Ptac*-ATS transposase gene ("derivative 2") comes from pNK2887, but the mini-transposon sequence and the backbone plasmid are replaced with a reconstructed one and pKP1588, respectively. The engineered transposon has a length of 2058 bp and contains *trp*-*lacZ*α gene fusion, the *kan* gene, and promoter sequence of the *araB* gene, bordered by inverted repeats of the outermost 70 bp of IS*10* Right (**Fig. 2**). The upstream stem structure of the *kan* gene was eliminated to avoid the interference with DNA sequence analysis. pKP1588 (T. Miki, unpublished) is a 7.8-kb mini-R–based cloning vector (derived from pRR12, a copy mutant of NR1) harboring *lacI*^{Q} and *cat* genes. As the donor plasmid pKP2371 has the *Ptac*-ATS (altered target specificity) transposase gene, the specificity of insertion is sufficiently low *(6)* to construct random-insertion library. Furthermore, the transposase gene is located on the donor plasmid outside the actual transposon, hence, no secondary transposition events are possible within recombinant phages or *E. coli* chromosome once the recombinant phages are transfected into recipient strain (away from pKP2371).

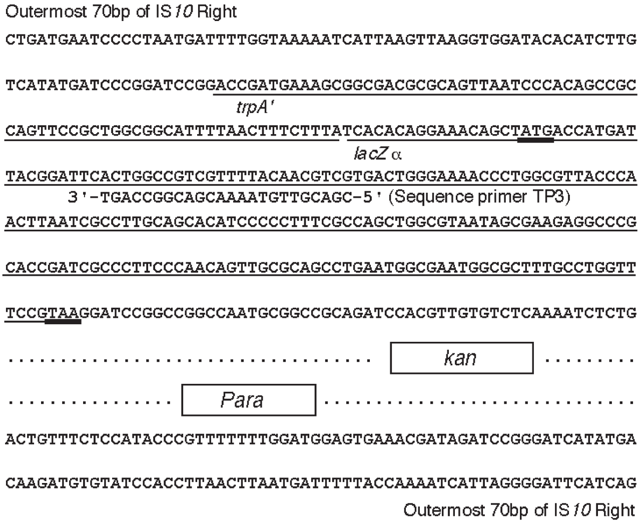

Outermost 70bp of IS10 Right
CTGATGAATCCCCTAATGATTTTGGTAAAAATCATTAAGTTAAGGTGGATACACATCTTG
TCATATGATCCCGGATCCGGACCGATGAAAGCGGCGACGCGCAGTTAATCCCACAGCCGC
trpA'
CAGTTCCGCTGGCGGCATTTTAACTTTCTTTATCACACAGGAAACAGCTATGACCATGAT
lacZ α
TACGGATTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCA
3'-TGACCGGCAGCAAAATGTTGCAGC-5' (Sequence primer TP3)
ACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAGCGAAGAGGCCCG
CACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAATGGCGAATGGCGCTTTGCCTGGTT
TCCGTAAGGATCCGGCCGGCCAATGCGGCCGCAGATCCACGTTGTGTCTCAAAATCTCTG
kan
Para
ACTGTTTCTCCATACCCGTTTTTTTGGATGGAGTGAAACGATAGATCCGGGATCATATGA
CAAGATGTGTATCCACCTTAACTTAATGATTTTTACCAAAATCATTAGGGGATTCATCAG
Outermost 70bp of IS10 Right

Fig. 2. The mini-Tn*10*(*lacZ*α-*kan*) transposon sequence and structure. For simplicity, not all of the 2058-bp sequence is shown. The *lacZ*α gene initiation and termination codons are shown by thick underline. Sequencing primer of TP3 is shown at the annealing site.

3. pKP2374 is a derivative of the mini-R plasmid pKP2305 harboring the *cIts857* gene of λ phage and the *sacB* gene of *B. subtilis* (*7*). pKP2305 (T. Miki, unpublished) is a NR1-based cloning vector with Ap^r marker. The temperature-sensitive repressor allows lysogenization of Kohara phages carrying a Km^r insertion mutation at 28°C and selection of Km^r haploid recombinants at 42.5°C. The *sacB* gene allows selection of haploid insertion mutants cured of pKP2374 by growing on TYCit sucrose plates.
4. LAM1 anneals 678 bp upstream of the *Bam*HI cloning site on the left arm of λ EMBL4, and LAM2 is located 640 bp downstream from the *Bam*HI cloning site on the right arm of EMBL4 (*8*).
5. Sequencing primer TP3 anneals 231 bp downstream from the 70 bp of IS*10* Right on mini-Tn*10*(*lacZ*α-*kan*) (**Fig. 2**).
6. Cloning of a variety of insertion mutations, which are carried, as a mixture, by mutagenized phage particles generated from a single Kohara library clone (described in **Sections 3.1.1**) was accomplished in this transduction step.
7. The prophage integrated in a "*cis*-diploid" mutant will excise if incubated at 42.5°C, as the *cI* repressor produced by pKP2374 is temperature-sensitive, and mutants would not form colonies as a result. On the other hand, Km^r haploid mutants (i.e., nonlysogenic derivatives of "*cis*-diploid" mutants) will grow at 42.5°C unless the disrupted gene is indispensable for growth. However, 42.5°C is too high for the excised λ phages to propagate efficiently. Hence, we chose 39°C as the incubation temperature in all experiments that needed excision and propagation of recombinant phages.
8. Amplification of long *E. coli* DNA fragments by PCR is a critical step. All components of the reaction are chilled before they are combined, and all operations are carried out at 0°C on an aluminum block placed on ice. DNase-free, RNase-free distilled water was used for preparation of reaction mixtures. Master mix for 96 samples was prepared by mixing carefully 768.6 μL sterile water, 210 μL Mg free 10× buffer, 197.4 μL 25 mM $MgCl_2$, 336 μL dNTP mixture (2.5 mM each), 21 μL Primer LAM1 (10 μM), 21 μL Primer LAM2 (10 μM) and 21 μL Taq polymerase (5 U/μL). Master mix (15 μL) was dispensed into each microwell of a reaction plate with 8-channel micropipettor, and 5 μL of phage suspensions were then added to each well. Mixing at this step is unnecessary. The wells were closed with MicroAmp Caps (8 Caps/strip), and the reaction mixtures were collected to bottom by centrifuging at 1000 rpm for 1 min. The plate was placed on a thermal cycler prewarmed to 95°C and incubated for 1 min for hot start. Then PCR program was started.
9. About 1% of cells lose pKP2374 per cell division when grown in the absence of selective pressure.
10. Because pKP2374 harbors *sacB*, only cells that have lost this plasmid can grow on a plate containing 5% sucrose.
11. The "*cis*-diploid" mutants that harbor both a wild-type allele and a mutated one in a chromosome are expected to be viable, irrespective of whether mutated gene is dispensable or indispensable. In contrast, "haploid" mutants harboring mutations in dispensable genes are viable, whereas "haploid" mutants harboring mutations in indispensable genes are lethal. Hence, by testing whether haploid insertion mutants arise from *cis*-diploid cells, *E. coli* genes indispensable for growth and survival can be identified.
12. Preliminary data will be obtained by replica-plating Km^r transductant plates described in **Section 3.1.2**, **step 10**, onto LBCit plates supplemented with kanamycin and ampicillin and incubating at 42.5°C. Results obtained in the experiments described in **Section 3.1.5**, **step 1**, will also be informative.

13. The frequency of formation of the nonlysogenic haploid derivatives of "*cis*-diploid" mutants largely depends on two factors: one is whether or not a disrupted gene is indispensable for *E. coli* growth, and the other is the location of an insert within the original Kohara clone relative to the ends of the cloned *E. coli* DNA fragment (the latter greatly influencing interchromosome recombination efficiency). Based on our control experiments with genes previously reported to be essential or nonessential (data not shown), we concluded that a disrupted gene can be deemed dispensable if the frequency of formation of nonlysogenic Km^R derivatives was higher than 10^{-4}. We asserted a gene as indispensable if this frequency was lower than 10^{-5} and the insertion was mapped farther than 300 bp from the nearest end of the cloned fragment in the Kohara clone. For more accurate judgment, the frequencies of nonlysogenic Km^R derivative formation can be measured in the presence of a complementing wild-type gene (provided on a plasmid *in trans*) and compared with those obtained in the absence of complementation.

Acknowledgments

We thank Dr. Y. Kohara for the Kohara bacteriophage λ clone miniset collection and Dr. H. Niki for encouragement. This work was supported in part by grants-in-aid for Scientific Research on Priority Areas from the Ministry of Education, Science and Culture of Japan and by CREST JST (Core Research for Evolutional Science and Technology, Japan Science and Technology Agency).

References

1. Kohara, Y., Akiyama, K., and Isono, K. (1987) The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50**, 495–508.
2. Jishage, M., and Ishihama, A. (1997) Variation in RNA polymerase sigma subunit composition within different stocks of *Escherichia coli* W3110. *J. Bacteriol.* **179**, 959–963.
3. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
4. Frschauf, A.-M., Lehrach, H., Polstka, A., and Murray, N. M. (1983) Lambda replacement vectors carrying polylinker sequences. *J. Mol. Biol.* **170**, 827–842.
5. Sanger, F., Niklen, S., and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467.
6. Kleckner, N., Bender, J., and Gottesman, S. (1991) Uses of transposons with emphasis on Tn*10*. *Methods Enzymol*. **204**, 139–180.
7. Ried, J. L., and Collmer, A. (1987) An *nptI-sacB-sacR* cartridge for constructing directed, unmarked mutations in gram-negative bacteria by marker exchange-eviction mutagenesis. *Gene* **57**, 239–246.
8. Ohshima, T., Aiba, H., Baba, T., Fujita, K., Hayashi, K., Honjo, A., et al. (1996) A 570-kb DNA sequence of the *Esherichia coli* K-12 genome corresponding to the 28.0–40.1 min region on the linkage map. *DNA Res.* **3**, 137–155.

14

High-Throughput Creation of a Whole-Genome Collection of Yeast Knockout Strains

Angela M. Chu and Ronald W. Davis

Summary

Gene disruption methods have proved to be a valuable tool for studying gene function in yeast. Gene replacement with a drug-resistant cassette renders the disruption strain selectable and is stable against reversion. Polymerase chain reaction–generated deletion cassettes are designed with homology sequences that flank the target gene. These deletion cassettes also contain unique "molecular bar code" sequence tags. Methods to generate these mutant strains are scalable and facile, allowing for the production of a collection of systematic disruptions across the *Saccharomyces cerevisiae* genome. The deletion strains can be studied individually or pooled together and assayed in parallel utilizing the sequence tags with microarray-based methods.

Key Words: gene disruption; homologous recombination; sequence tags; systematic disruption; yeast deletion; yeast knockout.

1. Introduction

When *Saccharomyces cerevisiae* became the first fully sequenced eukaryote, approximately 6000 open reading frames (ORFs) were identified (*1*). In itself, knowledge of the sequence did not directly translate into knowledge of gene functions, as 30% of this single-celled organism's gene functions are still not known (*2*). Yeast is a model organism that shares many of the same essential cellular processes as multicellular organisms, has both haploid and diploid cell types, and is tractable for genetic studies. A powerful method to elucidate gene function is gene disruption—removal of a functional protein allows for the study of its loss of function phenotype (*3*). Utilizing the yeast genome sequence, an international consortium of laboratories distributed the efforts to delete every ORF in the yeast genome. ORFs larger than 100 codons as well as "verified" shorter ORFs were disrupted from the translational start- to stop-codons and replaced with a kanamycin drug resistant marker and flanked by two unique 20-mer sequence tags that each serve to unequivocally identify each gene disruption (*4, 5*).

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

Well-established methods (6, 7) were optimized for high-throughput construction of the deletion strains utilizing the 96-well microtiter plate (MTP) format. ORF-specific polymerase chain reaction (PCR)-generated cassettes are integrated into the target locus with high efficiency through homologous recombination. ORF lists, genomic sequences, and ORF locations are obtainable from the Saccharomyces Genome Database (SGD; http://www.yeastgenome.org/) (8).

Two oligonucleotides are designed that specifically flank the start- and stop-codons for each ORF; appended to these sequences are unique 20-base sequence tags and common PCR primer sequences. The 3′ ends of the primers amplify the dominant drug marker for kanamycin (G418) resistance (9). An additional pair of oligonucleotides extends the cassette's yeast homology regions to 45 bases in a second round of PCR amplification. The deletion cassette is transformed into diploid yeast cells (10) and, through homologous recombination, replaces one copy of the wild-type gene (Fig. 1). Growth in the presence of G418 selects for recombinant colonies, which are verified for correct locus integration by the cassette with PCR (Fig. 2).

Heterozygous diploid deletion strains are subsequently sporulated; *MAT***a** and *MAT*α cell types are identified from the tetrads. A nonessential gene produces four viable haploid spores: two contain the intact ORF and two contain the disruption cassette. Essential gene deletions yield only two viable spores, each containing the undeleted wild-type ORF. Homozygous diploid strains are constructed from the mating of two independently isolated haploid mutants.

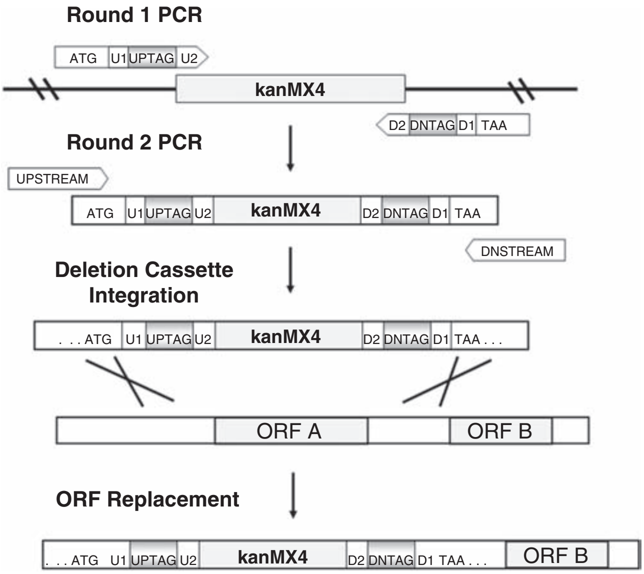
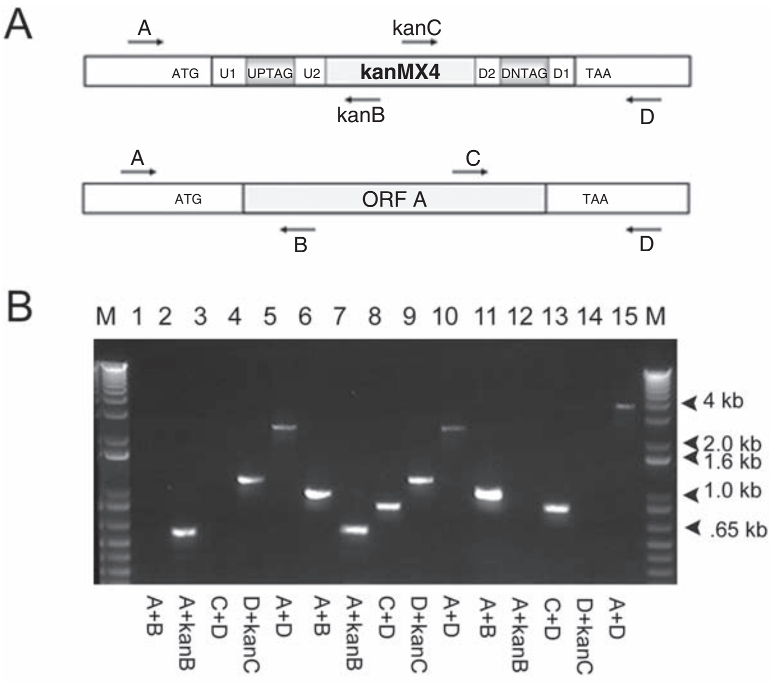

Round 1 PCR
ATG U1 UPTAG U2
kanMX4
D2 DNTAG D1 TAA
Round 2 PCR
UPSTREAM
ATG U1 UPTAG U2 kanMX4 D2 DNTAG D1 TAA
DNSTREAM
Deletion Cassette Integration
. . . ATG U1 UPTAG U2 kanMX4 D2 DNTAG D1 TAA . . .
ORF A
ORF B
ORF Replacement
. . . ATG U1 UPTAG U2 kanMX4 D2 DNTAG D1 TAA . . . ORF B

Fig. 1. Deletion cassette integration process. The first round of PCR uses the UPTAG and DOWNTAG primers to amplify the drug resistance marker KanMX. The second PCR reaction extends the homology of the deletion cassette to 45 bases upstream and downstream of the yeast ORF. Through homologous recombination of the end sequences, the deletion cassette module is accurately integrated into the chromosome, simultaneously removing and replacing the original ORF with the drug marker.



ΔYGL140C PCR sizes	A+B	A+kanB	C+D	D+kanC	A+D
deletion	--	576	--	1010	2244
wildtype	939	--	834	--	4320

Fig. 2. **(A)** Confirmation primer overview. **(B)** Confirmation PCR for yeast deletion *ygl140c::kanMX4*. Lanes 1 to 5 contain the haploid deletion, lanes 6 to 10 contain the heterozygous deletion, and lanes 11 to 15 contain wild-type yeast. Heterozygous deletion strains have both wild-type (A+B, C+D) and deletion (A+KanB, D+KanC) products (lanes 6 to 10). Haploid and homozygous diploid strains have the two deletion (A+KanB, D+KanC) products (lanes 2 and 4), and lack both wild-type (A+B, C+D) products (lanes 1 and 3). Lanes 5, 10, and 15 contain PCR products using the A+D primers.

Essential genes are only constructed in the heterozygous diploid form; a single copy of the wild-type gene is necessary to compensate for the loss of functional allele. Studies suggest that heterozygous loss of function phenotypes can be masked by the presence of the wild-type gene. However, deletion of a single copy of a nonessential gene in the diploid background can also lead to a reduced fitness level from the imbalance of gene product *(11)*. Thus, consortium efforts produced the yeast knockout (YKO) collection, which contains deletion strains in four different backgrounds: haploids of MAT**a** and MATα mating types, homozygous diploids (for nonessential genes), and heterozygous diploids *(12)* (**Note 1**). The collection is publicly available and is the foundation of many yeast genomic applications *(2)*.

2. Materials

2.1. Strains, Plasmids, and Oligonucleotides

1. BY4743: *MAT***a**/α *his3Δ1/his3Δ1 leu2Δ0/leu2Δ0 LYS2/lys2Δ0 MET15/met15Δ0 ura3Δ0/ura3Δ0*) (**Note 2**) (***10***) (American Type Culture Collection, Manassas, VA; ATCC 201390).
2. pFA6a-kanMX4: Plasmid containing the kanamycin (G418) resistance gene (**Note 3**).
3. Oligonucleotide primers (**Section 3.1**).

2.2. PCR Components

1. AmpliTaq Gold DNA polymerase enzyme (Applied Biosystems, Foster City, CA).
2. PCR buffer (50 mM Tris-HCl, pH 8.4, 250 mM KCl, 12.5 mM $MgCl_2$).
3. 20 mM deoxynucleotide triphosphates (dNTPs) solution.
4. 3 M sodium acetate, pH 5.2 (Sigma-Aldrich, St. Louis, MO).
5. Glycogen, 5 µg/mL (Ambion, Austin, TX).

2.3. Transformation Components

1. 1 M lithium acetate (LiAc) (lithium acetate dihydrate; Sigma-Aldrich). Prepare the solution using sterile ddH_2O; filter sterilize with an 0.22-µm filter. Store at room temperature.
2. 50% polyethylene glycol (PEG) (polyethylene glycol MW 3350; Sigma-Aldrich). Dissolve 15 g PEG-3350 in 16.5 mL sterile ddH_2O. Filter sterilize with an 0.22-µm filter. This is best made fresh but can be made in advance and stored at −20°C.
3. Lithium acetate + PEG solution (LiAc-PEG). Mix 12 mL 50% PEG with 1.5 mL 1 M lithium acetate and 1.5 mL sterile ddH_2O. Make this fresh before using.
4. Carrier DNA (10 mg/mL) (deoxyribonucleic acid sodium salt from salmon testes; Sigma-Aldrich). Dissolve salmon sperm DNA in sterile ddH_2O using a magnetic stirrer. Draw the mixture back and forth several times through a syringe fitted with a 15-gauge needle, then again with a 25-gauge needle. Mixture will be viscous. Aliquot into 1-mL tubes and store at −20°C. Before using, boil for 5 min and immediately place on ice.
5. Dimethyl sulfoxide (DMSO). Filter sterilize with an 0.22-µm filter. Diluted DMSO is made v/v with sterile ddH_2O or media and filter sterilized.

2.4. Media for Growth, Selection, and Maintenance

1. G418: Dissolve G418 disulfate (Sigma-Aldrich) to a final concentration of 50 mg/mL in sterile ddH_2O. Filter sterilize with an 0.22-µm filter and store at 4°C, protected from light. The activity of G418 varies from 400 to 800 µg per mg of total dry weight, so adjust accordingly for the final working concentration.
2. 40% dextrose: Dissolve 40 g dextrose w/v 100 mL ddH_2O. Autoclave and store at room temperature.
3. YPD plates: Mix 10 g yeast extract, 20 g bacto-peptone, and 20 g bacto-agar to 950 mL ddH_2O in a 2-L flask (**Note 4**). Autoclave and add 50 mL 40% dextrose. Cool to approximately 60°C before pouring into plates. For liquid media, omit agar.
4. YPD+G418 plates. Make YPD plate media, cool to approximately 60°C, and add 4 mL of G418 stock (final concentration 200 µg/mL). Mix thoroughly before pouring plates.
5. SD plates: Mix 6.7 g yeast nitrogen base without amino acids (BD, Franklin Lakes, NJ) and 20 g bacto-agar in 950 mL ddH_2O. Autoclave and add 50 mL 40% dextrose. Cool to 60°C before pouring plates.

6. Amino-acid supplements to add to SD plates: 10 amino acids make up the "complete" supplemental mix for SDC plates. Per 1 L media: 20 mg adenine, 20 mg arginine, 20 mg histidine, 30 mg leucine, 30 mg lysine, 20 mg methionine, 50 mg phenylalanine, 200 mg threonine, 20 mg tryptophan, 30 mg tyrosine, and 20 mg uracil (Sigma-Aldrich). Dissolve amino acid powders into 100 mL 65°C sterile ddH_2O, filter sterilize with an 0.22-µm filter, and mix with cooled media before pouring plates. To make dropout supplements, omit the appropriate amino acid(s) from the mix and proceed as above.
7. SDC-met and SDC-lys plates: Make SD plate media in 850 mL ddH_2O. Add in 50 mL 40% dextrose and 100 mL of the dropout supplement mix (**Section 2.4**, **item 6**), omitting either methionine or lysine, respectively.
8. Glycerol: Glycerol is diluted v/v with sterile ddH_2O and autoclaved. Store at room temperature.
9. Sporulation medium: Autoclave 1% potassium acetate, 0.005% zinc acetate w/v 950 mL ddH_2O. Supplement with 50 mL complete amino acid mix (**Section 2.4**, **item 6**). Store at room temperature.
10. Zymolase (Zymo Research Corp., Orange, CA).
11. 1 M sorbitol: Mix 18.2 g D-sorbitol (Sigma-Aldrich) w/v 100 mL ddH_2O. Sterilize by autoclaving and store at room temperature.

2.5. Accessories

1. Multichannel pipettors: 8- and 12-channel pipettors.
2. Reagent reservoirs (Matrix Technologies Corp., Hudson, NH).
3. OmniTrays (plates) (Nalge Nunc Intl., Rochester, NY).
4. 96-well microtiter plates (MTP) (Nunc).
5. 96-pin tool replicator (V-P Scientific, Inc., San Diego, CA).
6. 96-pin tool registrar device (Library Copier VP381, Colony Copier VP380; V-P Scientific, Inc.).
7. Foil sealing tapes for microtiter plates and brayer (Thermowell Sealing Tapes 6570, Corning Inc., Corning, NY).
8. Replica blocks: Round replica blocks fit 100-mm or 150-mm Petri plates (Cora Styles Needles ′N Blocks, Hendersonville, NC). A replica device for OmniTrays can be fashioned from a $2^7/_8 \times 4^1/_2$-inch block of nonporous materials (e.g., acrylic or aluminum) with three of the four corners filed off. Velvets can be held in place with a collar fashioned from an acrylic or a rubber gasket.
9. Velveteen pads (LabScientific, Inc., Livingston, NJ): Sterilize by autoclaving.
10. Glass beads (glass beads 5 mm; Sigma-Aldrich): Rinse with ddH_2O and store in a 250-mL Erlenmeyer screw-cap flask. Sterilize by autoclaving.
11. Vacuset 8-channel: This 8-channel vacuum aspiration device uses disposable pipette tips (Inotech Biosystems Intl., Inc., Rockville, MD).
12. Microtiter plate shaker incubator (VorTemp56 S2056; Denville Scientific, Inc., South Plainfield, NJ).
13. Micromanipulator system for yeast (**Note 18**).

3. Methods

3.1. Deletion Cassette Primer Design

The UPTAG and DNTAG primers are 74-mer oligonucleotides that consist of four components, from 5′ to 3′: 18 bases of sequence flanking the ORF including the start- (or stop-) codon, a common PCR priming site (U1 or D1; **Note 5**), a unique 20-bp sequence tag (**Note 6**), and sequence homologous to the kanamycin drug resistant marker (U2 or D2, **Note 5**).

A second set of overlapping primers extend the yeast homology regions further, improving gene targeting (**Note 7**). UPSTREAM and DNSTREAM primers are designed from the 45 bases of sequence immediately adjacent to the ORF and include the 18 bases of sequence of the UPTAG or DNTAG primers.

3.1.1. Confirmation Primer Design

ORF lists, genomic sequences, and ORF locations are obtained from the Saccharomyces Genome Database (SGD). PRIMER3 (**Note 8**) is used to pick the four confirmation primers, designated A, B, C, and D. Variables specific to the primer design are GC content (30% to 70%), Tm (59°C to 61°C), primer length (18 to 25 bp), target start = 200, and target length = 200. The ORF-flanking A and D primers are positioned 200 to 400 bp from the start- and stop-codons, respectively. The B and C primers are located within the coding region such that when used with the exterior primers, it results in PCR amplicons in the range of 250 to 1000 bp. The primers are then screened with MegaBlast (**Note 8**) to ensure that the primer sequences are unique within the genome. The two confirmation primers within the KanMX4 module, KanB and KanC (**Note 9),** when used with the A and D primers, respectively, give PCR products of 350 to 1000 bp in size (**Fig. 2A**).

3.2. Deletion Cassette Construction

3.2.1. Round 1: TAG Integration PCR

1. Resuspend the UPTAG and DNTAG primers to a final concentration of 10 pmol/µL, keeping the 96-well microtiter format. Using a multichannel pipettor, dispense 2 µL of each UPTAG and DNTAG primer into a 96-well PCR plate.
2. Mix 105 reactions' worth of PCR cocktail in a sterile solution reservoir. (The final concentrations or amounts for a single PCR reaction are shown in parentheses.)

105 Reactions:	1 Reaction:	
420 µL	4 µL	5× PCR buffer
21 µL	0.20 µL	20 mM dNTPs (0.2 µM)
21 µL	0.20 µL	5 U/µL Taq polymerase (1 unit)
1.05 µL	0.01 µL	20 µg/µL pFA6a-kanMX4 plasmid (~20 ng)
—	2 µL	10 pmol/µL UPTAG primer (1 pmol)
—	2 µL	10 pmol/µL DNTAG primer (1 pmol)
1217 µL H_2O	11.59 µL H_2O	
	20 µL total volume	

3. Transfer 16 µL of the master mix into the PCR plate with each of the UP/DNTAG primers.
4. PCR cycle conditions:
   10 min, 95°C
   30 s, 94°C |
   30 s, 55°C | × 22 cycles
   60 s, 72°C |

5 min, 72°C
Hold at 4°C

5. Check for amplification by loading 3 µL onto a 1% agarose gel; the cassette is ~1.6 kb in size.

3.2.2. Round 2: Homology Extension PCR

1. Resuspend the UPSTREAM and DNSTREAM primers to a final concentration of 10 pmol/µL, keeping the 96-well MTP format. Dispense into a 96-well PCR plate 4 µL of each UPSTREAM and DNSTREAM primer and 1 µL of the **Round 1** PCR product.
2. In a sterile solution reservoir, mix 105 reactions' worth of PCR cocktail. (Final concentrations for a single PCR reaction are shown in parentheses.)

105 Reactions:	1 Reaction:	
2100 µL	20 µL	5× PCR buffer
105 µL	1 µL	20 mM dNTPs (0.2 µM)
105 µL	1 µL	5 U/µL Taq Gold Polymerase (5 units)
—	1 µL	**Round 1** PCR product
—	4 µL	10 pmol/µL UPSTREAM primer (0.4 pmol each)
—	4 µL	10 pmol/µL DNSTREAM primer (0.4 pmol each)
7245 µL H_2O	69 µL H_2O	
	100 µL total volume	

3. Transfer 91 µL of the master mix into the PCR plate with each of the UP/DNTAG primers.
4. Cycle conditions are the same as for the **Round 1** PCR (**Section 3.2.1**, **step 4**). Check the **Round 2** amplification on a gel; the extended cassette size is ~1.7 kb.

3.3. PCR Precipitation

Before transformation, precipitate the PCR reaction to reduce its volume.

1. Transfer 95 µL PCR product to a 96-well microtiter plate.
2. Using a multichannel pipettor, add 10 µL NaAc + glycogen (for 105 reactions, mix 105 µL 5 µg/µL glycogen and 945 µL 3 M sodium acetate, pH 5.2), then add 100 µL isopropanol to each well.
3. Chill at −20°C for 20 min.
4. Centrifuge 10 min at 3000 rpm (~1600 × g).
5. Remove the supernatant with the Vacuset, being careful not to dislodge the pellet.
6. Wash with 100 µL chilled 70% ethanol and repeat the centrifugation step. Carefully remove the ethanol and air dry the pellets for 10 min. The plate can be stored at −20°C, sealed, at this step.

3.4. Transformation Protocol

Day 1: Inoculate 5 mL of YPD from a fresh BY4743 colony and grow overnight at 30°C.

Day 2:

1. Check the OD_{600} of the cells. BY4743 has an OD_{600} of 1 equal to 2×10^7 cells/mL.
2. Make 1 : 50 dilution into 10 mL YPD and grow at 30°C with shaking.

3. Check the OD_{600} at 3 to 4h and again at 6 to 8h. The density of the latter measurement should not be higher than 3 OD.
4. From the calculated growth rate (BY4743 has a 90-min doubling time), dilute 250mL cells to allow for a final OD of 1.5 to 2 with overnight growth (**Note 10**).

Day 3:

1. Check the culture's OD_{600}, which should be between 1.5 and 2.
2. Pellet cells by spinning at 3000rpm (~1600 ×*g*) for 5min.
3. Remove the supernatant and transfer the cells to 50-mL Falcon tubes.
4. Rinse the cells twice by resuspending in 50mL of 100mM LiAc, centrifuging 5min, and then removing the supernatant.
5. Resuspend the cell pellet in 1/100 of the total culture density of 100mM LiAc.
6. Add 1/9 of the cell volume of carrier DNA (**Note 11**).
7. Place the cells into a sterile reservoir and, using a multichannel pipettor, aliquot 25µL cells onto the PCR products.
8. Incubate at 30°C for 15min.
9. Add 125µL of the LiAc-PEG solution and mix by pipetting.
10. Incubate at 30°C for 30min.
11. Add 25µL DMSO to each well and mix by gentle pipetting.
12. Heat-shock the transformation plate at 42°C for 10min.
13. Pellet the cells by centrifuging the plate for 2min at 1500 ×*g*.
14. Carefully remove the LiAc-PEG mixture carefully without disturbing the cell pellet.
15. Add 200µL of YPD to each well and mix gently to dislodge the cell pellet (**Note 12**).
16. Incubate for 3h, shaking at 30°C.
17. Plate all the contents from one well onto one YPD+G418 plate by tipping 10 to 20 sterile glass beads onto each plate. Shake the plates back and forth to spread the cells with the beads; multiple plates can be stacked and shaken simultaneously. Remove the beads by tipping the beads into a beaker for reuse.
18. Incubate the plates at 30°C.

Day 5: Colonies appear within 2 to 4 days. Expect to see anywhere from a dozen to a hundred colonies. Pick four to eight healthy transformants and streak them out to single colonies onto fresh YPD+G418 plates. There are often colonies of differing sizes on a transformation plate. Avoid tiny colonies as these usually do not grow after streaking to the next round of selection.

3.4.1. Organization of Transformant Screening Plates

Organizing the isolates in MTP plates allows for storage of the colonies and prepares them for the subsequent confirmation steps.

1. Fill the wells of a MTP with 100µL YPD.
2. Pick seven colonies from the first transformation (A1) into column 1 (A1 to G1) of plate 1. The colonies from the second transformation (A2) continue in the next column (A2 to G2), and so on, until the last transformation (H12) fills column A12 to G12 of plate 8. Picking seven transformants allows for the eighth row to be used for controls.

3. Stamp the cells onto fresh YPD+G418 Omni plates using a sterile 96-pin tool.
4. Grow 2 days at 30°C.

3.4.2. Freezer Storage of the Transformant Plates

1. Add 7 µL of filter-sterilized DMSO to each well.
2. Seal the plates with foil sealing tapes, freeze, and store at −80°C.

3.4.3. Pinning from Microtiter Plates

1. Carefully peel off the sealing tapes while the MTPs are still frozen and let thaw on a flat surface.
2. Using a sterile 96-pin tool, carefully stir the cells.
3. Lift the 96-pin tool straight up without touching neighboring wells, checking that the pins are not dripping liquid into neighboring wells.
4. Pin cells onto agar plates by matching the guide pins of the replicator to the alignment holes of the 96-pin tool registrar.

3.4.4. Sterilizing the 96-Pin Tool

1. Set up three trays: sterile water, 70% ethanol, and 95% ethanol. The level of each should submerge the replicator pin deeper than the depth of the cell stocks (**Note 13**).
2. Place the replicator into the water for a minute to remove the yeast from the pins.
3. Transfer the replicator to the 70% ethanol; little or no cells should come off the pins.
4. Move the replicator through the 95% ethanol.
5. Gently tap the edge of the pins against the tray to remove excess ethanol, then flame.
6. Let the replicator cool completely before reuse.

3.5. Confirming Cassette Integration

3.5.1. Genomic DNA Preparation

1. Dispense 100 µL of the lysis cocktail (final concentration: 0.2 U/µL zymolase, 0.45% Tween 20, 0.45% NP-40, 50 mM KCl, 10 mM Tris pH 7.5, 1.5 mM $MgCl_2$) into the wells of a 96-well PCR plate.
2. Using a 96-pin tool, lightly touch the colonies on the agar plate and dip the 96-pin tool into wells. The liquid will be cloudy.
3. Cover the plate and place into the PCR machine. Incubate at 37°C for 60 min, then 100°C for 10 min.
4. Spin plate down in a centrifuge for 5 min to pellet the cells.
5. Use the supernatant as the DNA template in the PCR reactions.

3.5.2. Confirmation PCR

The following PCR protocol screens one MTP. It is useful to perform the corresponding wild-type and deletion reactions at the same time so that the PCR products can be visualized simultaneously (**Note 14**).

1. Using a multichannel pipettor, add 8 µL of each confirmation primer into each well of the top row of a 96-well PCR plate.
2. In a sterile basin, prepare the PCR reaction mix:

105 Reactions:	1 Reaction:	
420 µL	4 µL	5× PCR buffer
21 µL	0.20 µL	20 mM dNTPs (0.2 µM)
21 µL	0.20 µL	5 U/µL Taq polymerase (1 unit)
—	2 µL	genomic cell prep DNA
—	1 µL	20 pmol/µL forward primer (1 pmol)
—	1 µL	20 pmol/µL reverse primer (1 pmol)
1218 µL H_2O	11.6 µL H_2O	
	20 µL total volume	

3. Aliquot 136 µL of the PCR mix into each well of the top row.
4. Divide the PCR reaction mix by pipetting 18 µL into each of the seven rows.
5. Transfer 2 µL of the cell prep DNA into the appropriate wells.
6. PCR cycle conditions:
   10 min, 95°C (initial denaturation)
   30 s, 94°C |
   30 s, 57°C | × 30 cycles
   60 s, 72°C |
   5 min, 72°C (final elongation)
   Hold at 4°C
7. Load the PCR reactions on a gel and check for presence or absence of appropriately sized products.
8. Test for each of the A+B, A+kanB, C+D, and D+kanC amplicons; it is helpful to load the wild-type and deletion PCR products in alternating wells in the agarose gel.

Heterozygous deletion strains have both wild-type (A+B, C+D) and deletion (A+KanB, D+KanC) products (**Fig. 2B**). Haploid and homozygous deletions test positive for the A+KanB and D+KanC junctions but lack the wild-type (A+B, C+D) regions (**Notes 15** and **16**).

3.6. Haploid Deletion Strain Generation

From the heterozygous deletion strains, the two mating types are isolated along with the auxotrophic markers as set by the project: BY4741: *MAT***a** *his3Δ1 leu2Δ0 met15Δ0 ura3Δ0* and BY4742 *MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*. Each haploid has a different auxotrophic marker, further distinguishing between strain types; diploid strains are heterozygous for both. Generation of haploid strains and identification of essential genes is accomplished through dissection of the heterozygous diploid tetrads. Complementation methods are used to differentiate between the two mating types and verify retention of the auxotrophic markers.

3.6.1. Sporulation Protocol

1. Patch diploid deletion transformants onto a fresh YPD plate.
2. Grow 2 days.
2. Transfer a small patch of cells into 3 mL of sporulation media.
3. Grow, shaking at 22°C for 3 to 5 days. Check for tetrad formation under the microscope.
4. Dissect the tetrads onto YPD plates as per the lab's tetrad dissection setup (**Notes 17** and **18**).

5. Score tetrads for colony growth after 3 days.
6. Pick the tetrad colonies into 100 µL YPD into MTPs, following the dissection plate pattern for ease of identification. This plate will be used for further MTP screening and can be frozen down as described in **Section 3.4.1**.

3.6.2. Identification of Essential Genes

When the tetrad segregation is 2 : 2 (viable : lethal) on YPD, the gene deleted is essential for viability. Neither of the two remaining spores will be viable on G418 media as these tetrads contain the wild-type gene only. Save two independent copies of the heterozygous diploid isolates; this is an essential gene knockout (**Note 19**).

3.6.3. Isolating Haploid Deletion Strains

1. Using a 96-well 96-pin tool, pin the colonies from **Section 3.6.1** onto YPD (Fig. 3A) and YPD+G418 media.
2. Add controls (**Note 20**) onto the edge of the plate.
3. Grow at 30°C 1 to 2 days and score for G418-resistant strains (**Fig. 3B**).

3.6.4. Mating Protocol

Mating strain types can be identified by mating to the opposite mating type lacking different auxotrophic markers. Complementation through mating results in prototropic diploids that grow on minimal media.

Day 1: Patch BY4710 and BY4711 (**Note 21**) onto two 150-mm YPD plates so that the whole surface of the plate will grow into an even lawn.

Day 2:

1. Place a clean velvet pad on the replicator block and press the BY4710 plate onto the velvet, rotating the plate a few times to get an even lawn on the velvet. Press up to six YPD plates onto the velvet.
2. Repeat with BY4711, using a fresh velvet.
3. Change the velvet and gently press the deletion strain plate (**Section 3.6.3**) onto the velvet. Remove the strain plate and invert the BY4710 mating plate onto the velvet, making sure that the cells transfer onto the lawn.
4. Change the velvet and repeat, mating the deletion strains to BY4711.
5. Grow overnight at 30°C. The plates will grow into a solid lawn.
6. Replica plate the two mating plates onto SD plates and grow overnight at 30°C.
7. Score for growth on the plates and identify the mating types (**Fig. 3C, D**).

3.6.5. Selection for Auxotrophic Markers

In keeping with the collection genotypes, diploid deletion strains should grow on SDC-met-lys. Each haploid deletion mating type is associated with a separate auxotrophic marker: *MAT***a**'s and *MAT*α's grow on SDC-lys and SDC-met, respectively, and are nonviable when grown on SDC-met-lys media.

Testing for auxotrophic markers:

1. Place a sterile velvet pad on the replicator and invert the strain plate onto the velvet. Press gently to leave an impression of cells on the velvet.

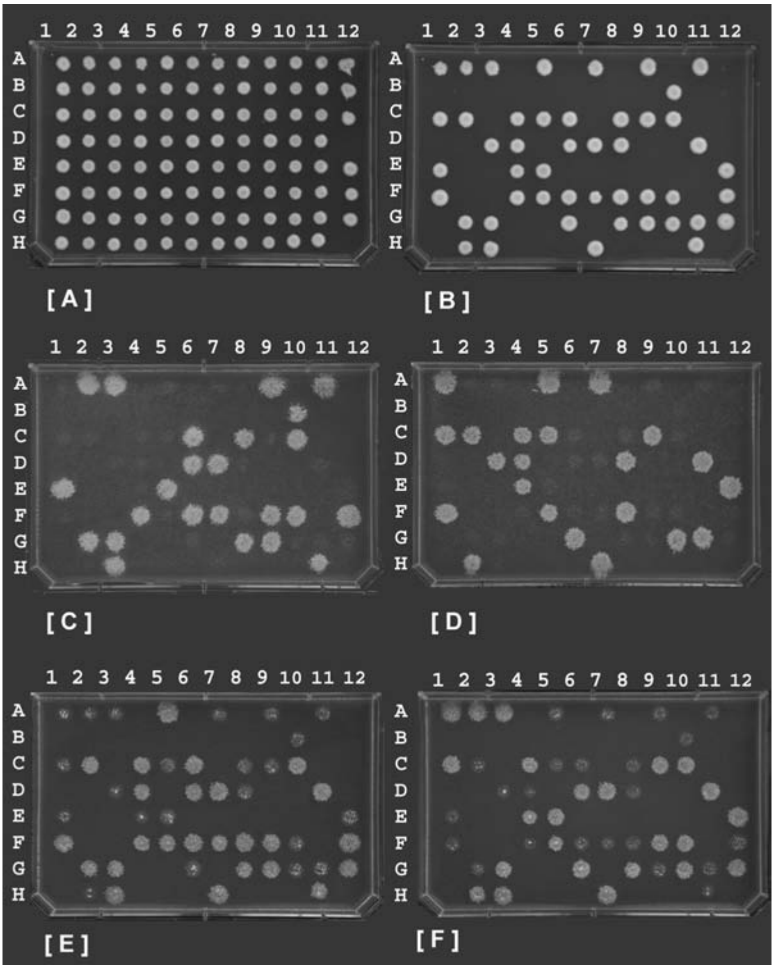


Fig. 3. Example of the selection process for drug resistance, mating type, and auxotrophic markers by replica plating across one microtiter plate. Wells A1 through D11 and wells E1 through H11 correspond with the 11 tetrad dissections (tetrads are arranged horizontally, the spores, vertically) from two different colonies of the same transformation. Column 12 contains control strains. A12, B12, and C12 are strains BY4741, BY4742, BY4743. E12, F12, and G12 are BY4741, BY4742, BY4743 containing the *ydl227c*::*KanMX* deletion. **(A)** Strains grown on YPD. **(B)** Strains pinned onto YPD+G418 media. Note the 2 : 2 segregation of the G418 resistant strains compared with **(A)**. **(C)** G418 resistant strains grown on SD media after mating with BY4710 (*MAT***a**). **(D)** G418 resistant strains grown on SD media after mating with BY4711 (*MAT*α). **(E)** Replica plating of plate B grown on SDC-met media. **(F)** Replica plating of plate B grown on SDC-lys media. C9 and C6, and H2 and F4 are pairs of haploid deletion strains from each transformant that test positively for each haploid mating type, *MAT***a** and *MAT*α, respectively, and their corresponding auxotrophic markers. This plate has multiple strains that fit the strain criteria; two pairs were chosen for this example.

2. Remove the plate and invert a SDC-met-lys plate onto the velvet.
3. Repeat for SDC-met and SDC-lys plates (**Note 22**).
4. Grow the plates overnight.
5. Score for growth (**Fig. 3E, F**).

3.7. Construction of the Homozygous Diploid Strains

1. Array the two haploid collections into 96-well plates, maintaining the original row-column designations.
2. Pin each haploid collection, using the plate register, onto YPD+G481 plates and grow overnight. Check that all strains grow.
3. Using the 96-pin tool and plate registrar, pin cells from the *MAT***a** deletions plate to a fresh YPD plate.
4. Sterilize the 96-pin tool.
5. Use the 96-pin tool to pin the cells from the *MAT*α deletions plate directly on top of the *MAT***a** strains by using the same alignment holes on the registrar.
6. Grow for 2 days at 30°C.
7. Streak each patch on YPD+G418 for single colonies and pick two to four isolates.
8. Verify that each strain is homozygous for the deletion cassette by PCR (**Section 3.5.2**), contains the appropriate auxotrophic markers by growth on SDC-met-lys (**Section 3.6.5**), is diploid by testing for lack of mating (**Section 3.6.4**), and undergoes sporulation (**Section 3.6.1**).

3.8. Long-Term Storage and Maintenance

3.8.1. Archiving Deletion Strains

Freeze down individual tubes of verified constructed deletion strain.

1. Patch the deletion strains onto YPD+G418 plates in ~2-cm-diameter patches (9 strains fit per 100-mm plate).
2. Grow 2 days.
3. Collect the cells by scraping up each patch with a sterile wooden toothpick. Resuspend the cells into a cryovial containing 1 mL YPD+15% glycerol. Store at −80°C.

3.8.2. Formatting the Finished Collection into 96-Well Plates

1. Patch the deletion strains onto YPD+G418 plates in ~2-cm-diameter patches (9 strains fit per 100-mm plate).
2. Grow 2 days.
3. Fill a deep-well MTP with 300 μL YPD per well.
4. Collect each patch of strains with a sterile wooden toothpick and place the cells into the corresponding well.
5. Seal the plate with a foil sealer and resuspend the cells (**Note 23**).
6. Aliquot 100 μL into 3 × 96-well plates that contain 100 μL 2× freezing media (**Note 24**).
7. Seal with a plate sealer and freeze.

3.8.3. Maintenance of Agar Plate Stocks

1. Pin strains from freezer stocks onto YPD+G418 using the registrar.
2. Grow 2 days at 30°C.
3. Transfer the strains from selection media to YPD by transferring the cells with a 96-well 96-pin tool (**Note 25**).

4. Grow for 2 days at 30°C.
5. Store at 4°C.

3.8.4. Making Deletion Strain Pools

1. Pin three copies of the deletion collection onto YPD+G418 plates.
2. Grow for 2 days (**Note 26**).
3. Harvest by using cell scrapers (BD Falcon) and resuspend cells in YPD+1× freezing media (**Note 24**), using 3 to 5 mL per plate.
4. Measure OD. Adjust the final OD to OD = 20.
5. Dispense into 100-µL aliquots and store at −80°C.

Notes

1. The YKO collection is available as individual strains or in sets, which can be obtained from American Type Culture Collection (http://www.atcc.org), Invitrogen (https://www.invitrogen.com), OpenBiosystems (http://www.openbiosystems.com), and EUROSCARF (http://web.uni-frankfurt.de/fb15/mikro/euroscarf/col_index.html). Project details can be found at: http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html.
2. BY4743 is a cross of strains: BY4741: *MAT***a** *his3Δ1 leu2Δ0 met15Δ0 ura3Δ0* and BY4742: *MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0* (American Type Culture Collection, 201388, 201389).
3. For the pFA6a-kanMX4 plasmid, contact peter.philippsen@unibas.ch *(8)*. Different antibiotic-resistant deletion marker plasmids are also available as a set at: http://web.unifrankfurt.de/fb15/mikro/euroscarf/data/Del_plas.html.
4. To facilitate mixing agar containing media, add a magnetic stir bar to the flask before autoclaving.
5. Common primer sequences (5′-3′): U1, GATGTCCACGAGGTCTCT; D1, CGGTGTCG GTCTCGTAG; U2, CGTACGCTGCAGGTCGAC; and D2, ATCGATGAATTCGAG CTCG.
6. The 20-mer sequence tags are from Affymetrix's *TAG3* microarray chip design (http://www.affymetrix.com).
7. Current oligonucleotide synthesis technologies produce quality extended-length primers and do not necessitate the use of two sets of construction primers for cassette production. We used primers synthesized on an Automated Multiplex Oligonucleotide Synthesizer (A.M.O.S.) in 5 to 10 nM amounts, organized in 96-well MTP format using standard phosphoamidite chemistry. When targeting highly homologous regions, such as gene families, flanking primers were extended in length (up to 90 bases) to improve gene targeting. For approximately 4% of the yeast genome, primers could not be chosen using the project's primer selection parameters. Primers used in the project can be found at: http://www-sequence.stanford.edu/group/yeast_deletion_project/Deletion_primers_PCR_sizes.txt.
8. Description and software downloads can be found at: http://frodo.wi.mit.edu/primer3/primer3_code.html for PRIMER3 *(13)* and http://www.ncbi.nlm.nih.gov/blast/ for MegaBlast *(14)*.
9. Kanamycin cassette primer sequences (5′-3′): KanB, CTGCAGCGAGGAGCCGTAAT; and KanC, TGATTTTGATGACGAGCGTAAT. Alternate primer sequences are KanB1, TGTACGGGCGACAGTCACAT and KanC3, CCTCGACATCATCTGCCCAGAT, which are located distally from the KanB and KanC sequences. The full sequence of pFA6a-KanMX4 can be accessed with accession numbers gi:2623975 and ASAJ2680 in Genbank or AJ002680.1 in EMBL-Bank.

10. To calculate the doubling time: $(T_{final} - T_{initial})/(1.44 \times \ln[OD_{final}/OD_{initial}])$, where $T_{final} - T_{initial}$ is the number of hours between two time points, and OD_{final} and $OD_{initial}$ are the optical density (OD_{600}) readings from the corresponding time points. Hours are defined as numbers, for example, use 1.5 h for a 90-min doubling time. To calculate the dilution amounts for an overnight culture, use:
    $(V \times [Y/X])/2^{(T/DT)}$, where V = volume of the culture in mL, Y = target OD, X = current OD, T = hours of growth, and DT = doubling time.
11. For example, if the final OD_{600} is 1.5 and the culture volume = 250 mL, resuspend the pellet in 3.75 mL 100 mM LiAc and add 417 µL of carrier DNA.
12. This is a good time to move the cells and media into a deeper, 1- to 2-mL volume 96-well microtiter plate before the shaking incubation. The Vortemp incubator is good for microtiter plates as the rotation diameter is only 5 mm. Otherwise, transfer the cells and attach the deep-well plate to a rotary platform with tape.
13. Tip box tops work well for rinsing the 96-pin tool. Change the water in the first tray often—every 5 to 10 pin cleanings.
14. For the control reactions, use wild-type DNA controls made from separate DNA preparations rather than from the cell preps, as this helps troubleshoot problems between the PCR reaction (enzyme, primers, or dNTP) versus DNA isolation problems.
15. The A+D primer pair can be used in lieu of one of the A+kanB or D+kanC reactions. In cases where the wild-type and deletion A+D PCR products are similar in size, the product can be digested with *HindIII* to check for the restriction site within the KanMX module.
16. For small ORFs (<100 amino acids), it is possible that all four confirmation primers are located outside the coding region. When this occurs, rather than looking for the absence of wild-type sized bands, the A+B and/or C+D would increase in size in the deletion strains.
17. Asci sac digestion: Pellet 100 µL of the sporulation suspension. Resuspend in 50 µL of 1 M sorbitol and 10 U zymolase and incubate at 30°C for 10 min. Add 150 µL sterile ddH_2O and immediately place on ice.
18. A comprehensive discussion of dissection scopes and methods can be found in: Sherman, F. Getting started with yeast (2002); modified from *Methods Enzymol.* **350**, 3–41: *see* http://dbb.urmc.rochester.edu/labs/sherman_f/startedyeast.pdf.
19. Because nonessential genes are made in multiple strain backgrounds, two independent copies of the essentials genes, designated isolates "A" and "B," are produced for duplication within the collection. A list of the essential genes in the YKO collection can be found at: http://www-sequence.stanford.edu/group/yeast_deletion_project/Essential_ORFs.txt.
20. Use controls of both wild-type and G418 resistant strains in haploid and diploid backgrounds.
21. BY4710 (*MAT***a** *trp1Δ63*) and BY4711 (*MATα trp1Δ3*) were used to test for complementation of mating types for the YKO collection (American Type Culture Collection; 200873, 200874).
22. Test haploids for SDC-met, SDC-lys, and SDC-met-lys as the two markers segregate separately; it is possible for a strain to be both -met and -lys.
23. The Vortemp works very well for resuspending cells; otherwise, seal and vortex carefully or use a multichannel pipettor for this step.
24. 2× freezing media is YPD v/v 14% DMSO or 30% glycerol. In high-density format, DMSO as a freezing agent is less viscous and easier to pin from. This method makes high-density stock plates with replicates.
25. For long-term storage and use from agar plates, it is not recommended to keep the deletion strains on G418 selection as this can lead to loss of heterozygosity (LOH) based on our observations.

26. Score plates for missing strains and slow or poor growth. These will be represented in lower quantities in the collection and may need to be grown supplemented back into the pool.

Acknowledgments

The authors would like to thank D.R. Richards, R.P. St. Onge, J.W. Mulholland, J. Horecka, and J.A. Brown for helpful discussion with the manuscript. Special thanks to L. Ni, L. Riles, S. Dow, and S. Soohai-Mahadeo for invaluable advice and discussion throughout the deletion project.

References

1. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., et al. (1996) Life with 6000 genes. *Science* **274**, 546–567.
2. Suter, B., Auerbach, D., and Stagljar, I. (2006) Yeast-based functional genomics and proteomics technologies: the first 15 years and beyond. *Biotechniques* **40**, 625–642.
3. Guthrie, C., and Fink, G., eds. (1991) *Guide to Yeast Genetics and Molecular Biology* (Methods Enzymology, Vol 194). San Diego: Academic Press.
4. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M., and Davis, R. W. (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat. Genet*. **14**, 450–456.
5. Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., et al. (1999) Functional characterization of the *Saccharomyces cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906.
6. Gietz, R. D., and Woods, R. A. (1994) High efficiency transformation with lithium acetate. In: Johnston, J. R., ed. *Molecular Genetics of Yeast, A Practical Approach*. Oxford: IRL Press, pp. 121–134.
7. Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F., and Cullin, C. (1993) A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic Acids Res*. **21**, 3329–3330.
8. Hong, E. L., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., et al. Saccharomyces Genome Database. Available at ftp://ftp.yeastgenome.org/yeast/.
9. Wach, A., Brachat, A., Pohlmann, R., and Philippsen, P. (1994) New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*. *Yeast* **10**, 1793–1808.
10. Brachmann, C. B., Davies, A., Cost, G. J., Caputo, E., Li, J., Hieter, P., and Boeke, J. D. (1998) Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* **14**, 115–132.
11. Deutschbauer, A. M., Jaramillo, D. F., Proctor, M., Kumm, J., Hillenmeyer, M. E., Davis, R. W., et al. (2005) Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* **169**, 1915–1925.
12. Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391.
13. Rozen, S., and Skaletsky, H. J. (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz, S. and Misener, S, eds. *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Totowa, NJ: Humana, pp. 365–386.
14. Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J Comput. Biol.* **7**, 203–214.

15

Analysis of Genetic Interactions on a Genome-Wide Scale in Budding Yeast: Diploid-Based Synthetic Lethality Analysis by Microarray

Pamela B. Meluh, Xuewen Pan, Daniel S. Yuan, Carol Tiffany, Ou Chen, Sharon Sookhai-Mahadeo, Xiaoling Wang, Brian D. Peyser, Rafael Irizarry, Forrest A. Spencer, and Jef D. Boeke

Summary

Comprehensive collections of open reading frame (ORF) deletion mutant strains exist for the budding yeast *Saccharomyces cerevisiae*. With great prescience, these strains were designed with short molecular bar codes or TAGs that uniquely mark each deletion allele, flanked by shared priming sequences. These features have enabled researchers to handle yeast mutant collections as complex pools of ~6000 strains. The presence of any individual mutant within a pool can be assessed indirectly by measuring the relative abundance of its corresponding TAG(s) in genomic DNA prepared from the pool. This is readily accomplished by wholesale polymerase chain reaction (PCR) amplification of the TAGs using fluorescent oligonucleotide primers that recognize the common flanking sequences, followed by hybridization of the labeled PCR products to a TAG oligonucleotide microarray. Here we describe a method—diploid-based synthetic lethality analysis by microarray (dSLAM)—whereby such pools can be manipulated to rapidly construct and assess the fitness of 6000 double-mutant strains in a single experiment. Analysis of double-mutant strains is of growing importance in defining the spectrum of essential cellular functionalities and in understanding how these functionalities interrelate.

Key Words: genetic interaction; molecular barcode; oligonucleotide microarray; SLAM; synthetic lethality; yeast knock-out strains.

1. Introduction

As evidenced by this and other volumes (*1, 2*), a major goal of the postgenomic era is to define the minimum set of functionalities required for robust "life" at both the cellular and organismal level, and beyond this, to understand the networks and pathways that weave these functionalities together in a way that provides both stability and adaptability to that life. Arguably, these goals can be achieved most efficiently in model systems where stable mutants are easily obtained, the genome can be facilely manipulated using molecular genetic techniques, and well-established, readily implemented

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

biochemical and/or phenotypic assays abound. One such model system is the budding yeast *Saccharomyces cerevisiae*.

Budding yeast has been the subject of classical and molecular genetic studies for decades, and these studies have produced fundamental discoveries about many key cellular processes including central metabolism, the secretory pathway, the cell cycle, signal transduction pathways, chromosome structure, replication and segregation, and transcriptional regulation. These discoveries are broadly applicable because yeast shares many functional homologues with other organisms, including humans. For these reasons, budding yeast has often been called the universal eukaryotic cell.

The budding yeast genome was the first to be sequenced in its entirety (3, 4). Soon thereafter, the coordinated efforts of several labs (collectively the *Saccharomyces* Genome Deletion Project) led to the generation of four yeast knock-out (YKO) strain collections: a heterozygous deletion diploid collection for ~6000 genes including ~1100 essential genes, as well as *MAT***a** and *MAT***alpha** deletion haploid collections, and a homozygous deletion diploid collection for nonessential genes (5–7). For this, open reading frames (ORFs; i.e., from ATG to STOP codon, inclusively), corresponding to known genes as well as with those inferred from the genomic sequence, were systematically replaced via homologous recombination with the *kanMX4* cassette (**Fig. 1**) (8), which confers dominant resistance to the antibiotic G418. Importantly, although all YKO deletion alleles carry an identical *kanMX4* selectable marker, each individual allele is uniquely "tagged" with two 20-bp DNA sequences that flank the *kanMX4* cassette proper. These UP and DOWN (or DN) TAGs function as molecular bar codes that specifically mark each deletion allele. Thus, the identity of individual YKO strains can be rapidly determined by polymerase chain reaction (PCR) amplification of the UPTAG and/or DNTAG from genomic DNA and subsequent sequencing of the PCR products. To facilitate this, all UPTAGs are flanked by universal priming sequences called U1 and U2; likewise, all DNTAGs are flanked by common sequences D1 and D2. Importantly, the ability to amplify any UPTAG or DNTAG with the same primer pairs (e.g., U1+U2c or D1+D2c, respectively) has also enabled researchers to handle YKO collections as single entities or pools. In this case, the presence and relative representation of individual strains within a population of all YKO strains can be evaluated by simultaneous PCR amplification of all UPTAGs and/or DNTAGs using fluorescently labeled primers, followed by hybridization of the resultant mixture of PCR products to an UPTAG and DNTAG oligonucleotide microarray (5, 6, 9, 10).

The *S. cerevisiae* YKO collections and the genome sequence, on which they are based, have been invaluable resources, allowing researchers to study gene expression, protein localization, protein-protein interaction, and gene function on a global scale. The YKO collections continue to grow as new data implicate previously overlooked small open reading frames (i.e., those less than 100 codons) as authentic genes (11–13). The YKO collections have defined the spectrum of eukaryotic genes individually essential or important for cellular life. However, it is clear from many focused genetic studies that essential genes do not define all essential functions. Rather, two or more genes or pathways often mediate certain critical cellular functions. In some cases, such functional redundancy or genetic buffering is effected by truly homologous factors. For example, many eukaryotes have multiple copies of each histone gene. More frequently

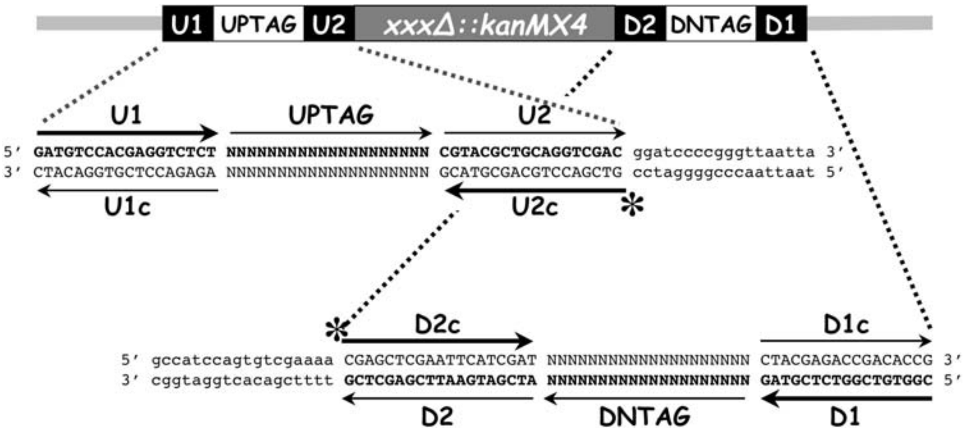

U1
UPTAG
U2
xxxΔ::kanMX4
D2
DNTAG
D1
U1
UPTAG
U2
5′ GATGTCCACGAGGTCTCT NNNNNNNNNNNNNNNNNNNN CGTACGCTGCAGGTCGAC ggatccccgggttaatta 3′
3′ CTACAGGTGCTCCAGAGA NNNNNNNNNNNNNNNNNNNN GCATGCGACGTCCAGCTG cctaggggcccaattaat 5′
U1c
U2c
*
*
D2c
D1c
5′ gccatccagtgtcgaaaa CGAGCTCGAATTCATCGAT NNNNNNNNNNNNNNNNNNNN CTACGAGACCGACACCG 3′
3′ cggtaggtcacagctttt GCTCGAGCTTAAGTAGCTA NNNNNNNNNNNNNNNNNNNN GATGCTCTGGCTGTGGC 5′
D2
DNTAG
D1

Fig. 1. Generic *xxx*$_N$Δ::*kanMX4* YKO allele with sequence details. For each YKO allele, the relevant open reading frame has been precisely replaced via homologous recombination with a *kanMX4* cassette that confers resistance to the antibiotic G418 (*5, 6*). The *kanMX4* cassette consists of the *kanr* open reading frame of the *E. coli* transposon Tn903 fused to transcriptional and translational control sequences of the *TEF* gene of the filamentous fungus *Ashbya gossypii* (*8*). These sequences are flanked by short UPTAG and DNTAG sequences that are unique for each gene. The TAG sequences themselves are flanked by short universal sequences that can be used as priming sites to PCR amplify the TAGs (*5, 6*). Thick gray lines denote genomic sequence flanking the *xxx*$_N$Δ::*kanMX4* deletion allele where *XXX* is any yeast gene. Positions and sequences of the universal priming sites are indicated. UPTAG and DNTAG sequences (N_{20}) function as probes on the Hopkins TAG Array (*22*). A fluorescently labeled extract suitable for hybridization to the Hopkins TAG Array is prepared by PCR amplification using genomic DNA prepared from a complex pool of YKO mutants as the template. Cy3- or Cy5-labeled UPTAGs are amplified using U1 and Cy-labeled U2c (asterisk) oligonucleotides as primers. Cy3- or Cy5-labeled DNTAGs are amplified using D1 and Cy-labeled D2c (asterisk) oligonucleotides as primers. Here, "c" simply indicates a complementary sequence. Importantly, in the PCR reactions, the Cy-labeled primer is present in a 10-fold molar excess over its unlabeled counterpart. This allows for preferential amplification of Cy-labeled strands during later PCR cycles. These Cy-labeled strands are complementary to probe sequences on the microarray. Prior to hybridization, PCR reactions are denatured, then allowed to anneal with an appropriate blocking oligonucleotide mixture—U1+U2 for UPTAG PCRs or D1+D2 for DNTAG PCRs. This step is meant to reduce spurious hybridization.

and usually more difficult to predict, functional redundancy is effected by molecularly distinct pathways that culminate in similar or compensatory outcomes. For example, pathways for homologous recombination and nonhomologous end-joining collaborate to maintain genome integrity by repairing double-strand breaks albeit by different mechanisms. Thus, to fully describe the minimal requirements for eukaryotic life, one needs to comprehensively assess the fitness or viability of strains containing pairwise or higher-order combinations of mutations. This is a daunting task, even in a simple model organism like budding yeast, where a thorough study of double mutants based on classical approaches would involve at least 25 million genetic crosses!

We have developed a simple approach to rapidly and systematically generate double-deletion mutant yeast strains by transformation and to assess their viability (*10, 14–16*). This approach, called dSLAM, or diploid-based synthetic lethality analyzed by microarray, takes advantage of the fact discussed above that YKO strains can be manipulated as pools and that the presence of any particular YKO strain in a population can be monitored by PCR amplification of the UPTAGs and DNTAGs followed by hybridization to an oligonucleotide microarray (**Fig. 2**). For dSLAM, a pool of all heterozygous deletion diploids is transformed *en masse* with a single query gene disruption construct after which single- and double-mutant haploid pools are derived by sporulation and differential selection. Representation of haploid YKO strains in each pool is then assessed by differential labeling of PCR-amplified TAGs followed by competitive hybridization to a single TAG microarray slide.

In so far as dSLAM seeks to assess the phenotypes of double mutants, it is complementary to two other approaches—namely, SGA (synthetic genetic array) (*17–19*) and

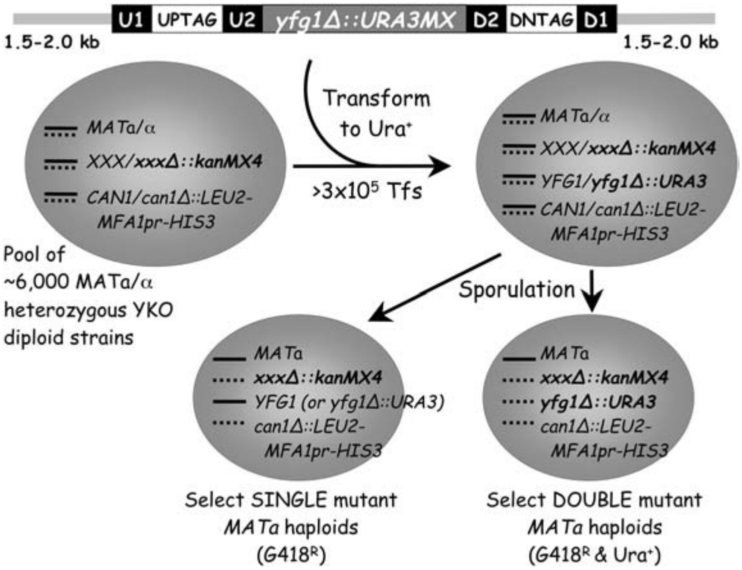


Fig. 2. Overview of dSLAM approach. Conceptually, dSLAM is straightforward (*10, 14, 16*). To start, one needs only two, albeit sophisticated, reagents: (1) a query gene disruption fragment marked by either *URA3* (shown here as *yfg1*Δ*::URA3MX*) or *NatMX* (*21*) and having at least 1.5 kb of query ORF flanking sequence on each side of the selectable marker; and (2) a comprehensive pool of heterozygous YKO deletion diploids that all carry the SGA reporter or "Magic Marker" (*can1*Δ*::LEU2-MFA1pr-HIS3*). The heterozygous YKO pool is transformed *en masse* with the query gene disruption fragment and the resultant pool of transformants is sporulated to generate haploid cells. Viable single-mutant *MAT***a** $xxx_N\Delta$*::kanMX4* spores that inherit the Magic Marker can be selected on SC+URA-LEU-HIS-ARG+CAN+G418 medium (MM+URA). Double-mutant *MAT***a** $xxx_N\Delta$*::kanMX4 yfg1*Δ*::URA3* spores that inherit the Magic Marker can be selected on SC-URA-LEU-HIS-ARG+CAN+G418 medium (MM-URA). The relative representation of UPTAGs and DNTAGs in these two haploid pools will vary according to whether a synthetic genetic interaction exists between the query gene disruption allele and a given $xxx_N\Delta$*::kanMX4* allele (target gene).

E-MAP (epistatic miniarray profile) (**20**)—that employ genetic crosses to construct double mutants from viable haploid YKO strains on a large, if not genome-wide, scale. Although experimental approaches based on microarrays have their own drawbacks, dSLAM is arguably superior to SGA and E-MAP in that dSLAM relies on heterozygous YKO strains as a starting point. Heterozygous diploid strains are less likely than their haploid counterparts to accumulate secondary mutations that modify the deletion phenotype. In addition, working with the complete heterozygous YKO collection allows one to monitor genetic interactions between a nonessential query gene and either nonessential target genes (i.e., synthetic lethality or synthetic fitness) or essential target genes (i.e., synthetic rescue or suppression). Finally, for the nonessential genes, every possible double-mutant strain can be constructed and characterized *in duplicate* with only 5000 dSLAM experiments, as opposed to 25 million crosses, making it possible to generate a comprehensive synthetic lethal data set within a relatively short period (we anticipate completing the first pass within the next 2 years). For these reasons, we have undertaken dSLAM on a genome-wide, high-throughput scale. Below, the protocols currently being used for this project are outlined.

2. Materials

1. *Heterozygous deletion diploid collection*. The heterozygous deletion diploid collection is available from several sources including Open Biosystems (Yeast Heterozygous Diploid, cat. no. YSC1071, Huntsville, AL); Invitrogen Corporation (Heterozygous Diploid A, cat. no. 95401.H4R3, Carlsbad, CA); and American Type Culture Collection (ATCC; Heterozygous diploid, cat. no. GSA-6, Manassas, VA). The complete set includes 5996 strains arrayed across sixty-seven 96-well plates numbered 201–263, 270–271, and 280–281. The overall genotype of the *Saccharomyces cerevisiae* heterozygous deletion collection is *MAT***a***/MAT***alpha** *ura3Δ0/ura3Δ0 leu2Δ0/leu2Δ0 his3Δ0/his3Δ0 met15Δ0/MET15 lys2Δ0/LYS2* $xxx_N\Delta$*::kanMX4/*XXX_N^+.
2. *Plasmid pXP346 containing the "Magic Marker" cassette*. Plasmid pXP346 carries the SGA reporter (**17–19**) or "Magic Marker" cassette (i.e., *can1 5′ UTR::LEU2-MFA1pr-HIS3::can1 3′ UTR*) (**10**). A 4.6-kb fragment suitable for transformation can be released by *Spe*I-*Pst*I restriction enzyme digestion of plasmid pXP346. At least 20 μg of the transforming linear DNA fragment are required per transformation to obtain a sufficient number of transformants. pXP346 transformants are selected on synthetic complete medium lacking leucine (SC-LEU) medium. Targeted integration of the Magic Marker at the *CAN1* locus confers recessive canavanine resistance.
3. *Query gene disruption fragment*. As with the Magic Marker cassette, at least 20 μg of the transforming linear query gene disruption fragment is required per transformation to obtain a sufficient number of transformants. For the protocol given below, the DNA should be in sterile dH$_2$O or Tris-EDTA buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA), in a total volume of 100 to 150 μL (**Note 1**; *see* Ref. **21**).
4. YEPD + G418 + carbenicillin solid medium (1% yeast extract, 2% bactopeptone, 2% dextrose, 2% agar, 1.5 mM L-tryptophan, 200 μg/mL G418, 100 μg/mL carbenicillin) in OmniTrays (Nalge Nunc International Corp, Rochester, NY).
5. YEPD liquid (1% yeast extract, 2% bactopeptone, 2% dextrose, 1.5 mM L-tryptophan).
6. SC-LEU plates, 150-mm and 100-mm diameter.
7. SC-URA-LEU plates, 150-mm and 100-mm diameter.
8. SC+URA-LEU-HIS-ARG+CAN+G418 plates ("MM+URA" plates; *see* Ref. **16** for detailed recipe), 150-mm and 100-mm diameter.

9. Synthetic complete medium lacking uracil and leucine (SC-URA-LEU)-HIS-ARG+CAN+G418 plates ("MM-URA" plates; *see* Ref. *16* for detailed recipe), 150-mm and 100-mm diameter.
10. Liquid sporulation medium (1% potassium acetate, 0.005% zinc acetate, 300 µM histidine-HCl).
11. Carbenicillin: Dissolve at 100 mg/mL in sterile water and filter sterilize. Use at 100 µg/mL final concentration. Store 1000× stock at 4°C.
12. L-Canavanine (Sigma-Aldrich, cat. no. C1625, St. Louis, MO): Dissolve at 60 mg/mL in sterile water and filter sterilize. Use at 60 µg/mL final concentration. Store 1000× stock at –20°C.
13. G418 (Geneticin; Invitrogen Corp., cat. no. 11811-031): Dissolve at 200 mg/mL effective concentration in sterile water and filter sterilize. Use at 200 µg/mL final concentration. Store 1000× stock at –20°C.
14. Sterile dH$_2$O.
15. Lithium acetate (LiOAc), 1 M and 0.1 M, sterile.
16. 50% polyethylene glycol (PEG-3350): Dissolve PEG in sterile water, adjust volume once air bubbles have disappeared, and filter sterilize. Store working stock at room temperature. Store any additional aliquots at –20°C for future use.
17. Sonicated Herring Sperm DNA, 10 mg/ml (Promega Corporation, cat. no. D1816, Madison, WI): Just before use, aliquot amount needed into separate microcentrifuge tube and heat denature DNA at 100°C for 5 min, then place on ice for at least 5 min.
18. DMSO (dimethyl sulfoxide; from Qbiogene [Irvine, CA] or Sigma-Aldrich).
19. 5 mM CaCl$_2$, sterile.
20. Glass beads, 3-mm diameter (Sigma, cat. no. 11-312A), autoclaved.
21. Glycerol, 80% and 15% in water, sterile.
22. Epicentre MasterPure Yeast DNA Purification Kit (Epicentre, cat. no. MPY80200, Madison, WI).
23. Isopropanol, 100%.
24. Ethanol, 70% and 100%.
25. RNAse A, 20 mg/mL (Invitrogen, cat. no. 12091-039).
26. QIAamp DNA Micro Kit (Qiagen, cat. no. 56304, Valencia, CA).
27. 3 M sodium acetate.
28. TE (10 mM Tris-HCl, pH 8.0; 1 mM EDTA).
29. Molecular biology grade dH$_2$O. Purchase commercially. Prepare and store 1 mL aliquots in a clean room setting.
30. Heat sealable bags (4 cm × 12 cm; Ampac Packaging, Cincinnati, OH).
31. Lidded compartment boxes (e.g., Alpha RHO Inc., cat. no. 776C-6-P, Fitchburg, MA).
32. 2× Ex Taq DNA Polymerase Premix (Takara Bio USA, cat. no. TAK RR003 Madison, WI).
33. 10 : 1 Cy3 UP, Cy5 UP, Cy3 DN, and Cy5 DN Fluorescent Primer Mixes (**Note 2**; *see* Ref. *23*).

    Prepare separate 10 µM stocks of the following unlabeled oligonucleotides in 10 mM Tris-HCl, pH 6.5:

U1	`5′-GATGTCCACGAGGTCTCT-3′`
D1	`5′-CGGTGTCGGTCTCGTAG-3′`

    Prepare separate 100 µM stocks of the following Cy3- or Cy5-labeled oligonucleotides in 10 mM Tris-HCl, pH 6.5:

Cy U2c	`5′ Cy*-GTCGACCTGCAGCGTACG-3′`
Cy D2c	`5′ Cy*-CGAGCTCGAATTCATCGAT-3′`

Mix together equal volumes of primers as specified below to generate four different primer mixes in which the fluorescent primer is at 50 µM final concentration and the unlabeled primer is at 5 µM final concentration. Aliquot 10 : 1 primer mixes into clearly labeled tubes and store working stocks in the dark at –20°C. For long-term storage, keep at –80°C.

10 : 1 mix	Fluorescent primer	Unlabeled primer
Cy3 UP	Cy3 U2c	U1
Cy5 UP	Cy5 U2c	U1
Cy3 DN	Cy3 D2c	D1
Cy5 DN	Cy5 D2c	D1

34. *Blocking oligonucleotides*. Order each of the indicated blocking oligonucleotides on a 10 µmol scale, desalted, and resuspend in 10 mL high-purity water for a ~1 mM solution.
    (a) 0.5 mM UPTAG blocking oligonucleotide mix. Mix together equal volumes of the following then aliquot and store at –20°C:

1 mM	U1	5′ - GATGTCCACGAGGTCTCT - 3′
1 mM	U2-3	5′ - ACGCTGCAGGTCGAC - 3′

    (b) 0.5 mM DOWNTAG blocking oligonucleotide mix. Mix together equal volumes of the following then aliquot and store at –20°C:

1 mM	D1	5′-CGGTGTCGGTCTCGTAG-3′
1 mM	D2-3	5′-GATGAATTCGAGCTCG-3′

35. 10% Triton X-100. Mix 90 mL high-purity water with 10 mL Ultrapure 100% Triton (USB, cat. no. 22686, Cleveland, OH) and filter through 0.2-µm filter.
36. 0.1 M dithiothreitol (DTT): Prepare a large volume, filter through 0.2-µm filter, and store as 1-mL and 10-mL aliquots at –20°C. Discard leftover DTT after it has been thawed once (or reserve for use during prehybridization).
37. *Hybridization buffer*:

	For 1 liter	
1 M NaCl	200 mL	5 M NaCl (USB, cat. no. 75888)
100 mM Tris-Cl, pH 7.5	100 mL	1 M Tris-HCl, pH 7.5 (USB, cat. no. 22639)
0.5% Triton X-100	50 mL	10% Triton X-100
	650 mL	High-purity water

    Add DTT to a final concentration of 1 mM to the requisite volume of hybridization buffer just before use (e.g., 1/100 volume of 0.1 M DTT stock).
38. *Hopkins TAG Array or other bar-code microarray.* For dSLAM, as well as for other types of functional profiling of YKO populations, fluorescently labeled PCR-amplified UPTAGs and DNTAGs are hybridized to a YKO bar-code oligonucleotide microarray. We currently use the "Hopkins ***TAG*** Array" designed by Daniel Yuan (*see* Ref. ***22*** and **Note 3** for details; *see also* Ref. ***24***) and manufactured by Agilent Technologies (Yeast Barcode Oligo Microarray, cat. no. G2518A Option 006, Santa Clara, CA).
39. 20× SSPE, 0.2 µm filtered (USB, cat. no. 75890). *Note*: 20× SSPE solution consists of 3 M NaCl, 200 mM sodium phosphate, and 20 mM EDTA, pH 7.4. Use high-purity water to make 6× SSPE and 0.06× SSPE dilutions.

40. *Hyb wash buffer I*:

	For 50 mL	
6× SSPE	50 mL	6× SSPE
0.05% Triton X-100	250 µL	10% Triton X-100
1 mM DTT	500 µL	0.1 M DTT

41. *Hyb wash buffer II*:

	For 50 mL	
6× SSPE	50 mL	0.06× SSPE
1 mM DTT	500 µL	0.1 M DTT

See **Note 4.**

42. *Stripping buffer*:

	For 50 mL	
1% SDS	2.5 mL	20% SDS (USB, cat. no. 75832)
10 mM EDTA	1.0 mL	0.5 M EDTA (USB, cat. no. 15694)

3. Methods

3.1. General Strategy for dSLAM

For dSLAM experiments, a representative pool of ~6000 heterozygous deletion diploids is first prepared from the commercially available yeast heterozygous YKO diploid strain collection. The heterozygous *xxx*$_N$Δ*::kanMX4/XXX*$_N^+$ diploid pool is then transformed *en masse* with the *can1*Δ*::LEU2-MFA1pr-HIS3* cassette ***(10)*** originally developed by Tong et al. for SGA ***(17, 18)***. The SGA cassette or so-called Magic Marker allows for direct selection of *MAT***a** haploid strains after sporulation of the heterozygous deletion diploid pool (see below). The resultant Magic Marker heterozygous deletion diploid pool is then transformed with a query gene disruption construct (e.g., *yfg1*Δ*::URA3*, where *YFG* means "your favorite gene"; **Fig. 2**). Targeted integration of this DNA to the *YFG1* locus via homologous recombination produces a pool of double heterozygous deletion diploids. Finally, the *yfg1*Δ*::URA3MX* transformed heterozygous deletion diploid pool is sporulated, and *MAT***a** *xxx*$_N$Δ*::kanMX4* haploids are selected via the Magic Marker in the presence or absence of uracil. The former population is the control pool in which all viable *xxx*$_N$Δ*::kanMX4* haploid strains (and their associated UPTAGs and DNTAGs) should be represented as single, if not double, mutants. The latter population is the experimental pool that should in theory contain only viable, relatively fast-growing *yfg1*Δ*::URA3MX xxx*$_N$Δ*::kanMX4* double-mutant haploid strains. Genomic DNA is prepared from each pool and UPTAGs and DNTAGs are PCR amplified from the single- and double-mutant pools with Cy5- and Cy3-labeled primers, respectively (**Fig. 3**). The fluorescent PCR products are then hybridized to a single oligonucleotide array. A high Cy5 : Cy3 ratio indicates deletion alleles that interact with the query gene disruption to produce a synthetic fitness defect or synthetic lethality.

3.2. Construction of a Heterozygous xxx$_N$Δ::kanMX4/XXX$_N^+$ Deletion Diploid Pool

1. To create a representative pool of heterozygous YKO strains, first spot the entire heterozygous deletion diploid collection (67 plates) onto solid YEPD + G418 + carbenicillin medium in OmniTrays using a sterile 96-pin bolt replicator (V&P Scientific, Inc., San Diego, CA) to generate ~5-mm patches after growth (**Note 5**).

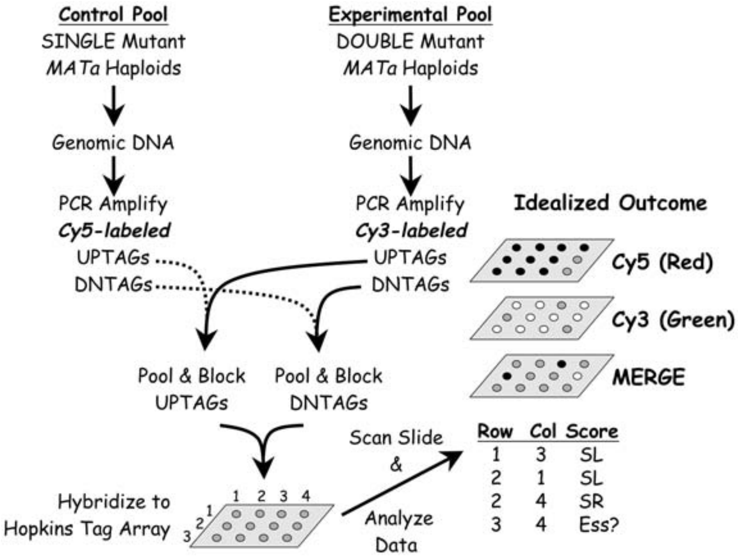


Fig. 3. Assessment of TAG representation in YKO pools. Representation of UPTAGs and DNTAGs in the single versus double YKO mutant haploid pools is assessed by PCR amplification of the TAGs using fluorescent primers and subsequent hybridization of differentially labeled PCR products to a TAG oligonucleotide microarray. Here, UPTAGs and DNTAGs present in the single-mutant pool are amplified using a Cy5-labeled primer (black hybridization signal), whereas UPTAGs and DNTAGs present in the double-mutant pool are amplified using a Cy3-labeled primer (white hybridization signal). The separate PCR reactions are mixed together and hybridized to a single microarray slide. Fluorescence data is gathered at two wavelengths for each feature on the microarray, and the ratio of Cy5/Cy3 intensities for each feature is determined. A high ratio indicates a potential synthetic lethal (SL) or synthetic fitness (SF) interaction as the particular YKO allele is underrepresented in the double-mutant pool. Such is the case for the genes corresponding with positions 1,3 and 2,1 on the hypothetical microarray. A low ratio might indicate a synthetic rescue (SR) interaction as the particular YKO strain appears to perform better when the query gene is disrupted. Such is the case for the gene corresponding with position 2,4 on the hypothetical microarray. Finally, some features will consistently show hybridization signals that are close to background such as the one at position 3,4. This result is expected, for example, if the YKO allele is lethal. (*See* **Note 20** for other explanations.) Methods for microarray data analysis have been well described elsewhere (*[10, 15, 22, 29, 30]*; *see* **Chapter 25**).

2. Incubate the OmniTrays at 30°C for 2 days.
3. Inspect the YEPD + G418 + carbenicillin plates for contamination and make note of any slow-growing mutants (e.g., many ribosome protein gene mutants exhibit haplo-insufficiency) or uneven inocula. Carefully scrape the patches from one of each type of plate into sterile 15% glycerol, combining all scrapes together to make a single homogeneous pool. Individual patches can be scraped from the "backup" OmniTray(s) and added to the pool to improve the representation of slower growing diploids.
4. Be sure the pool is thoroughly mixed. Adjust cell density to ~60 to 75 OD_{600} equivalents per milliliter and store 1- or 2-mL aliquots at −80°C.

3.3. Construction of a Haploid-Convertible Magic Marker Heterozygous $xxx_N\Delta$::kanMX4/XXX_N^+ Deletion Diploid Pool

In order to reliably obtain haploid pools after sporulation, the SGA reporter, affectionately called the Magic Marker, must be introduced into all heterozygous *$xxx_N\Delta$::kanMX4/XXX_N^+* deletion strains that compose the starting pool. This can be done simply by transforming the previously generated heterozygous YKO pool (**Section 3.2**) *en masse* with the *can1Δ::LEU2-MFA1pr-HIS3* reporter. For this, follow the high-efficiency integrative transformation protocol described below (**Section 3.4**). Set up several independent transformations in parallel, using 10 to 20 μg of *Spe*I-*Pst*I digested pXP346 plasmid per transformation. The bulk of each transformation should be plated onto one or several 150 mm × 25 mm plates containing solid SC-LEU medium for *LEU2* selection. Be sure to also determine the transformation efficiency of each individual transformation as described in **Section 3.4.3**, **step 9**. It is critical to obtain $\geq 5.0 \times 10^5$ independent Leu$^+$ transformants per transformation to avoid random loss of certain YKO mutants from the population. The Leu$^+$ transformants from all individual transformations that match this criterion can be harvested with filter-sterilized 15% glycerol, pooled, and stored as 1- or 2-mL aliquots at −80°C.

Alternatively, one can obtain our recently generated Magic Marker heterozygous YKO diploid collection from Open Biosystems or ATCC, then pool the individual Leu$^+$ strains that compose this collection in the same manner as described in **Section 3.2**. In fact up to now, we have used such a defined Magic Marker heterozygous diploid pool as the starting point for production dSLAM experiments. However, the occasional dSLAM user might find it simpler and more cost effective to make their own *en masse* Magic Marker transformed pool.

3.4. High-Efficiency Yeast Transformation of the "Magic Marker" $xxx_N\Delta$::kanMX4/XXX_N^+ Heterozygous Deletion Diploid Pool

3.4.1. Growth of Cells

1. Thaw one 1-mL aliquot of the Magic Marker *$xxx_N\Delta$::kanMX4/XXX_N^+* heterozygous deletion diploid pool on benchtop. Invert tube several times to resuspend cells.
2. Inoculate 0.5 mL into each of two 1-L flasks containing 250 mL YEPD liquid supplemented with 100 μg/mL carbenicillin. Be sure to reserve a small amount of YEPD liquid to use as a blank for following culture density. The starting OD_{600} should be ~0.15 ODU per milliliter.
3. Vigorously shake cultures at 30°C for 5–6 h, or until OD_{600} triples to ~0.5 ODU per milliliter. This will give ~250 ODU or 5×10^9 cells, enough for 10 dSLAM transformations.

3.4.2. Prepare Competent Cells for Transformation

1. Transfer cells to sterile Corning 250-mL polypropylene (PP) centrifuge tubes.
2. Harvest cells by centrifugation in a Sorvall RC-5C or comparable centrifuge (3000 rpm, 6 min, 22°C).
3. Decant medium. Loosen cell pellets by vortexing, then resuspend in sterile dH_2O to wash.
4. Combine all cells in one bottle at this stage and recentrifuge.
5. Decant liquid. Loosen pellet and resuspend in 0.1 M LiOAc to wash. Recentrifuge.

6. Decant liquid. Loosen pellet and resuspend yeast cells in 0.1 M LiOAc for ~10.5 mL final volume.

3.4.3. Transformation of Yeast Cells

1. Aliquot 1 mL competent yeast cells to each of 10 microcentrifuge tubes. Pellet cells by centrifugation at 2000 rpm for 30 s.
2. Flick away or aspirate *most, but not all*, of the 0.1 M LiOAc. Vortex or shake tubes to loosen cell pellets in the residual liquid, then store tubes at room temperature.
3. Prepare a cocktail containing the following amounts of reagents *per transformation*:
   620 µL 50% PEG-3350
   90 µL 1.0 M LiOAc
   40 µL 10 mg/mL herring sperm DNA (heat denatured and quick-chilled just before use)
4. Add 20 to 40 µg of query gene disruption fragment DNA (e.g., *yfg1Δ::URA3MX* PCR-amplified DNA) in a total volume of 100 to 150 µL to each ~100 µL cell aliquot. Vortex to *completely* resuspend cells and mix with the DNA.
5. As soon as possible, add 750 µL PEG cocktail to each transformation. Immediately invert microcentrifuge tube several times to premix. Once PEG cocktail has been added to all transformations, vortex or shake all samples vigorously for 30 to 60 s to thoroughly mix (using multitube shaker).
6. Incubate cells at 30°C for 30 min with gentle agitation (e.g., place tubes on a rocker or a roller drum).
7. After the 30°C incubation, add 100 µL DMSO to each transformation and gently mix.
8. Heat shock cells at 42°C for 14 min (using an aluminum heating block or water bath). Gently invert tubes once or twice during heat shock period to promote even heating. Centrifuge to pellet cells at 2000 rpm for 1 to 2 min.
9. Carefully aspirate the PEG/DMSO supernatant, switching tips for each sample.
10. Add 1 mL 5 mM $CaCl_2$ and resuspend cells thoroughly by gently pipetting up and down. The final volume should be ~1.1 to 1.2 mL.
11. Let cells recover in 5 mM $CaCl_2$ at room temperature for at least 5 min, *but not longer than 15 min*.
12. *Titer transformation efficiency.* Transfer 2 µL of well-resuspended transformed cells to a second microcentrifuge tube containing 198 µL 5 mM $CaCl_2$. Plate 100 µL and 10 µL of this 1 : 100 dilution onto 100-mm SC-URA-LEU plates. For the 10-µL aliquot, pipet 90 µL of 5 mM $CaCl_2$ onto plate first. Spread transformed cells evenly.
13. Plate the remainder of each transformation onto a 150-mm SC-URA-LEU plate. If plates are fresh (wet), it will be necessary to re-pellet the transformed cells at 2000 rpm for 30 s and remove ~500 µL supernatant. Resuspend cells in remaining liquid then transfer to the large SC-URA-LEU plate. Spread the cells evenly. The addition of ~20 to 25 sterile 3-mm glass beads to each plate can speed this process. Retain the glass beads in the lid after inverting the plates—they can be used later when harvesting the transformants.
14. Incubate all plates at 30°C for 2 full days (48 h).

3.4.4. Assess Transformation Efficiency and Harvest Transformants

1. Determine the number of colonies on the 100-mm SC-URA-LEU plates and calculate the total number of Ura^+ transformants. Also note colony size and morphology as occasionally heterozygosity at the query gene locus can affect growth. In practice, we usually obtain between 3×10^5 and 5×10^5 transformants, although even higher yields are possible with this protocol. The stability of the query gene disruption marker (e.g. *URA3*) in the primary transformants should be assessed (**Notes 6** and **7**; *see* Refs. 25–27).

2. Provided the desired number of stable transformants is obtained, scrape the lawn of yeast transformants from each 150-mm SC-URA-LEU plate using two additions of sterile water (7 mL, then ~3 to 5 mL). The cells can be dislodged easily by swirling ~20 to 25 sterile 3-mm glass beads on the surface of the plate along with the water. Avoiding the glass beads, transfer both "scrapes" to a single 15-mL conical tube containing 2.5 mL sterile 80% glycerol. Adjust the second addition of sterile water so as to recover a total of 10 mL of scraped cells. The final 12.5-mL volume corresponds with ~16% glycerol, a concentration suitable for storing yeast at –80°C. Mix well.
3. Determine the OD_{600} of each transformed pool. For this, mix 10 µL of cells with 990 µL sterile water in a 1-mL plastic cuvette (1 : 100 dilution).
4. Proceed to sporulation (see below) and/or freeze several 25-ODU aliquots at –80°C for future use (**Note 8**). For high-throughput dSLAM we store transformant pools in Screen-Mate Latch Racks (1.4-mL round-bottomed tubes; Matrix Technologies Corp., Hudson, NH).

3.5. Sporulation of the yfg1Δ::URA3MX/YFG1$^+$ xxx$_N$Δ::kanMX4/XXX$_N^+$ *Double Heterozygous Deletion Diploid Pool and Selection of Haploid Pools*

3.5.1. Outgrowth of Transformed Cells in Rich Medium

1. For each query gene transformation, inoculate 25 ODU of the *yfg1Δ::URA3MX/YFG1*$^+$ *xxx*$_N$*Δ:kanMX4/XXX*$_N^+$ heterozygous diploid pool into 50 mL YEPD (use a 250-mL flask).
2. Vigorously shake culture at 30°C for 2 to 3 h.
3. After 2 to 3 h, transfer culture to a sterile 50-mL conical tube. Harvest cells by centrifugation in a swinging bucket rotor (3000 rpm, 4 min, 22°C).
4. Decant supernatant, vortex tube to loosen cell pellet, then resuspend cells in 25-mL sterile water to wash. Recentrifuge to pellet cells, decant supernatant, and vortex tube to loosen cell pellet.

3.5.2. Sporulation of Heterozygous Diploid Pools

1. Resuspend outgrown and washed cells in 50-mL sporulation medium. Decant cells into a sterile 250-mL flask, ideally one with a screw cap to reduce evaporation.
2. Vigorously shake sporulating culture at 25°C for 5 days.
3. After 5 full days, transfer culture to a sterile 50-mL conical tube and, as needed, adjust volume up to 50 mL with fresh sporulation medium. This will simplify calculations later (see below).
4. Transfer 500 µL of the culture to a 1-mL cuvette containing 500 µL water to prepare a 1 : 1 dilution. Mix well, avoiding air bubbles, and measure OD_{600}.
5. Harvest cells by centrifugation, decant supernatant, and resuspend sporulated cells in sterile water at a final concentration of 10 ODU/mL (**Note 9**).

3.5.3. Assess Sporulation Efficiency

Place 10 µL of the sporulated culture on a microscope slide and assess sporulation efficiency microscopically. Score 100 to 200 cells. The frequency of sporulation-positive cells for the YKO genetic background under these conditions is usually 30% to 50%. If the frequency is outside this range, the amount of cells plated for haploid selection should be adjusted accordingly.

3.5.4. Select Single- and Double-Mutant Haploid Pools

Provided the diploid cultures have sporulated, they are next plated onto two types of media, both of which select for *MAT***a** haploids by virtue of the Magic Marker ("MM"). Neither parental His$^-$ Can^S diploids nor His$^-$ *MAT***alpha** haploid progeny should grow on these media. A "Control" or reference population of all viable *MAT***a** *xxx*$_N$*Δ::kanMX4* single-mutant haploids (as well as any viable *MAT***a** *xxx*$_N$*Δ::kanMX4 yfg1Δ::URA3MX* double-mutant haploids) are selected on "MM+URA" medium (SC+URA-LEU-HIS-ARG+CAN+G418). Thus, this pool should contain UPTAGs and DNTAGs corresponding with YKOs for all nonessential genes. The "Experimental" population of viable *MAT***a** *xxx*$_N$*Δ::kanMX4 yfg1Δ::URA3MX* double-mutant haploids are selected on "MM-URA" medium (SC-URA-LEU-HIS-ARG+CAN+G418). This pool should contain UPTAGs and DNTAGs only for those YKOs that are not lethal in a *yfg1Δ::URA3MX* background. In principle, greater discrimination between viable *MAT***a** *xxx*$_N$*Δ::kanMX YFG1*$^+$ single mutants and inviable *xxx*$_N$*Δ::kanMX4 yfg1Δ::URA3MX* double mutants can be obtained by selecting the Control population on "MM+URA+5-FOA" medium (SC+URA-LEU-HIS-ARG+CAN+G418+5-FOA) that permits growth only of Ura$^-$ *MAT***a** *xxx*$_N$*Δ::kanMX4 YFG1*$^+$ single mutants. However, in practice for high-throughput dSLAM, we routinely use MM+URA medium.

1. Evenly spread 0.4 mL (~4 ODU) of the 10 ODU/mL sporulated culture onto each of two 150-mm haploid selection plates: a MM+URA plate (Control) and a MM-URA plate (Experimental). Alternatively, inoculate 0.4 mL of the 10 ODU/mL sporulated culture into 100 mL each of liquid MM+URA and MM-URA medium.
2. Titer single and double *xxx*$_N$*Δ::kanMX4* mutants. To determine the frequency of G418^R *URA3*$^+$ double-mutant haploids and of G418^R haploids overall, prepare a 1 : 2000 dilution of each 10 ODU/mL sporulated culture. Typically, we first prepare a 1 : 100 dilution first, then dilute this 20-fold. Plate 100-µL aliquots of the 1 : 2000 dilution onto 100-mm MM-URA and MM+URA haploid selection plates and spread evenly.
3. Invert all plates and incubate at 30°C for 2 full days (48 h). A shorter incubation might be possible if using liquid haploid selection.

3.5.5. Assess Haploid Selection and Harvest Control and Experimental Pools

1. Count the colonies on each 100-mm plate, then multiply by 8000 to determine the approximate number of haploids selected on each of the large plates. Also note colony size and morphology, as growth phenotypes conferred by disruption of the query gene should be manifest at the haploid stage (**Note 10**).
2. Scrape lawns from the large MM+URA and MM-URA plates using ~ 10 mL sterile water per plate and glass beads as described above. Transfer each individual "scrape" to a 15-mL conical tube. There is no need to use a second addition of water here.
3. Determine the OD_{600} of each collected haploid pool. For this, mix 10 µL of cells with 990 µL H_2O in a 1-mL plastic cuvette (1 : 100 dilution).
4. For each sample, freeze several 25-ODU aliquots at –80°C. For high-throughput dSLAM, MM+URA and MM-URA haploid selectants are typically stored in separate Matrix ScreenMate Latch Racks (**Note 11**).
5. Finally, for those experiments in which a subset of primary Ura$^+$ transformants appeared 5-FOA^R **(Note 7)**, replica print the 100-mm MM-URA plate to MM+URA+5-FOA to assess the stability of the *URA3* marker. Usually, but not always, unstable transformants are lost during the sporulation process.

3.6. Preparation of High-Quality Genomic DNA from Haploid Pools

3.6.1. Overview

For each query gene used for a dSLAM transformation, high-quality genomic DNA (gDNA) is prepared from the matched Control and Experimental haploid pools (**Note 12**). For this purpose, we currently use two kits in sequence: a MasterPure Yeast DNA Purification Kit (Epicentre), followed by a QIAamp MinElute Column (Qiagen), although other approaches can be used *(16, 28)*. The resultant gDNA will be used as a template for wholesale PCR amplification of UPTAGs and DNTAGs using fluorescently labeled universal primers.

3.6.2. Cell Lysis and Precipitation of DNA Using the Epicentre MasterPure Yeast DNA Purification Kit

1. For a given query, thaw microcentrifuge tubes or latch-rack tubes containing the Experimental and Control haploid yeast pools at room temperature.
2. If cells are in a latch-rack tube, resuspend each population (~25 ODU) thoroughly by gentle pipetting and transfer as much as possible to a clean 1.5-mL microcentrifuge tube.
3. Pellet the yeast cells by centrifugation in a microcentrifuge at 3000 rpm for 2 to 5 min. Remove the supernatant by aspiration, changing tips for each sample, briefly re-spin, and aspirate residual liquid. Vortex tubes to loosen cell pellets.
4. The Epicentre MasterPure Yeast DNA Purification Kit is available in 10 and 200 purifications sizes. The 200 purifications kit (cat. no. MPY80200) contains (also see **Note 13**):

Yeast Cell Lysis Solution	60 mL
MPC Protein Precipitation Reagent	50 mL
RNase A @ 5 μg/μL	200 μL
1× TE Buffer	7 mL

5. *Lyse yeast cells.* Add 300 μL Epicentre Yeast Cell Lysis Solution to each loosened 25 ODU yeast cell pellet. Resuspend cells completely in the lysis solution by either vortexing or pipetting.
6. Incubate resuspended cells at 65°C for 15 min, then place tubes on ice for 5 min.
7. Add 150 μL MPC Protein Precipitation Reagent and vortex samples for 10 s. Then, pellet cellular debris by centrifugation in a microcentrifuge at ≥10,000 rpm for 10 min.
8. *Isopropanol precipitate the gDNA.* Transfer the supernatant to a clean microcentrifuge tube containing 500 μL isopropanol (not provided in kit). Mix thoroughly by inversion to precipitate the DNA. At this point, it is okay to let samples sit at room temperature for a short period of time.
9. Collect DNA (plus RNA) precipitate by centrifugation in a microcentrifuge at ≥10,000 rpm for 10 min. A white pellet should be readily visible at this point.
10. Remove the isopropanol supernatant with a pipette and discard. Wash the DNA pellet with 500 μL 70% ethanol. Re-spin at ≥10,000 rpm for 1 to 2 min if necessary.
11. Carefully remove the ethanol wash with a pipette and discard. Briefly re-spin for 1 min and remove any residual ethanol on the DNA pellet using a fine pipette tip. A vacuum aspirator can be used to remove the ethanol, provided a fresh pipette tip is used for each sample.
12. Allow DNA pellet to air-dry, then resuspend the sample in 95 μL TE buffer. Gentle vortexing and/or heating at 37°C can speed this process.

13. Once the sample is dissolved, add 5 µL 20 mg/mL RNAse A (Invitrogen; cat. no. 12091-039), mix well, and incubate at 37°C for 30 min to 1 h before proceeding with the QIAamp DNA Micro Kit.

3.6.3. gDNA Purification Using QIAamp DNA Micro Kit

1. The Qiagen QIAamp DNA Micro Kit is available in a 50 purifications size. The kit (cat. no. 56304) contains (**Note 14**):
   QIAamp MinElute Columns
   QIAGEN Proteinase K
   Carrier RNA
   Buffers AW1, AW2, and AE
   Collection tubes (2 mL)
2. For a total sample volume of 100 µL (see above), add 10 µL Qiagen Buffer AW1 to the resuspended DNA. The sample will become very cloudy.
3. Add 250 µL Qiagen Buffer AW2. Mix sample by vortexing for 10 s. The sample should clear somewhat and a crystalline precipitate will begin to form.
4. Transfer the entire volume including the crystalline precipitate (~360 µL) to a QIAamp MinElute Column that has been placed in a new clean 2-mL collection tube. Spin column at 8000 rpm for 1 min, and discard collection tube containing the flow-through.
5. Place the MinElute Column in a new, clean, 2-mL collection tube. Add 500 µL Qiagen Buffer AW2 to column, spin at 8000 rpm for 1 min, and discard collection tube containing the flow-through.
6. Place the MinElute Column in a new, clean, 2-mL collection tube or a 1.5-mL microcentrifuge tube. Centrifuge MinElute Column at 14,000 rpm for 3 min to dry the membrane.
7. Place the MinElute Column in a clean, 1.5-mL microcentrifuge tube and add 100 µL Qiagen Buffer AE to elute the genomic DNA. Be careful to direct the elution buffer to the center of the membrane, but avoid touching the membrane with the pipette tip. Incubate at room temperature for *at least* 1 min, then centrifuge at 14,000 rpm for 1 min. Preheating the elution buffer to 50°C can improve yield.
8. After centrifugation, the flow-through that has collected in the microcentrifuge tube contains the eluted gDNA. Discard the used MinElute Column.
9. *Ethanol precipitate the gDNA.* Add 10 µL of 3 M sodium acetate, pH 5.2 (i.e., 1/10 volume) to the ~100 µL eluted gDNA. Mix well. Add 200 µL cold 100% ethanol (i.e., 2 volumes) to the gDNA and invert the microcentrifuge tube several times to mix well. Sample should become cloudy. At this point, it is okay to let samples sit at room temperature for a short period of time.
10. Centrifuge sample at 14,000 rpm for 10 min to pellet the DNA precipitate. A DNA pellet, albeit smaller than before, should be visible at this point. Carefully remove supernatant and wash the DNA-containing pellet with 500 µL 70% ethanol as described above. Dry pellet at 37°C for 15 to 30 min or air dry at room temperature overnight.
11. Resuspend genomic DNA samples in 25 µL of 10 mM Tris-HCl, pH 8.0. Store the gDNA at −20°C or −80°C (**Note 15**).
12. *Determine gDNA concentration.* Prepare a 1 : 10 dilution of each gDNA sample and determine its DNA concentration. For this, we use either PicoGreen dsDNA Quantitation Reagent (Molecular Probes, Inc.), according to the manufacturer's instructions, or a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE). The quality of the gDNA can also be assessed by agarose gel electrophoresis. In theory, 25

ODU of haploid cells in G1 contains roughly 10 µg of genomic DNA. We typically recover at least 5 µg in 25 µL using this protocol.

3.7. Preparation of Cy5- and Cy3-Labeled Fluorescent Extracts Using gDNA Isolated from Matched Control and Experimental Haploid Pools

3.7.1. Overview

With this protocol, the molecular bar codes (i.e., UPTAGs and DNTAGs) that uniquely mark each *xxx*$_N$*Δ::kanMX4* (and *yfg1Δ::URA3MX*) deletion allele are amplified wholesale from gDNA prepared from either the Control (MM+URA) or the Experimental (MM-URA) haploid pool. Control gDNA is amplified with Cy5-labeled U2c or D2c primers and unlabeled U1 or D1 primers, respectively (**Fig. 1**). Experimental gDNA is amplified with Cy3-labeled U2c or D2c primers and unlabeled U1 or D1 primers. Importantly, the ratio of Cy-labeled U2c (or D2c) to unlabeled U1 (or D1) primers in each PCR reaction is 10:1. This leads to asymmetric PCR in later cycles and favors the generation of Cy-labeled single strands that are complementary to the TAG oligonucleotides on the Hopkins TAG Array. Finally, UPTAGs and DNTAGs are amplified in separate PCR reactions. Thus, for each dSLAM experiment there will be four PCR reactions:

Cy3-labeled Experimental Pool UPTAGs
Cy5-labeled Control Pool UPTAGs
Cy3-labeled Experimental Pool DNTAGs
Cy5-labeled Control Pool DNTAGs

3.7.2. Gather Reagents in Clean Area

1. Thaw 2× Ex Taq Premix and the four 10 : 1 Primer Mixes (Cy3 UP, Cy5 UP, Cy3 DN, and Cy5 DN) at a second “clean” bench or hood. Gently mix each tube and briefly spin in a dedicated microcentrifuge (i.e., one that has never been used for yeast cells, yeast gDNA, or yeast-derived PCR products).
2. Thaw gDNA samples to be amplified at the “clean” bench where gDNAs are normally prepared, mix well, and briefly spin to collect liquid. Transfer tubes to a “clean” rack and, if necessary, bring the rack to a location near, but not in, the “clean” area where dSLAM PCR reagents are handled. Be sure to change gloves after handling the gDNA samples and before reentering the second “clean” area.

3.7.3. Prepare a Master Mix for Each of the Four 10 : 1 Primer Mixes

The general recipe for a single reaction is specified below. Prepare enough Master Mix for the actual dSLAM experiments at hand, as well as for a “No DNA” control hybridization and other control hybridizations, as desired (**Section 3.7.5**).

Reagent	Single reaction
10 : 1 UP or DN Primer Mix (50 µM : 5 µM)	6 µL
dH_2O	23 µL
2× Ex Taq Premix	30 µL
(gDNA)	(1-2 µL)
Final volume	~60 µL

To prepare the Master Mixes, first add molecular biology grade dH_2O to each of four prelabeled microcentrifuge tubes (i.e., Cy3 UP, Cy5 UP, Cy3 DN, Cy5 DN) using a new tip for each aliquot. Next, add the appropriate 10:1 Primer Mix. Finally, add 2× Ex Taq Premix to each Primer Mix tube. Mix gently but well, then briefly spin in the dedicated microcentrifuge.

3.7.4. Set Up PCR Reactions

1. For each pair of Control and Experimental samples, place four 0.2-mL PCR tubes or one 8-tube PCR strip in a prechilled rack. Aliquot 59 µL of each Master Mix to a separate PCR tube in each set. Add the *last* aliquot of each Master Mix to the "No DNA" negative control tube in case there is not enough Master Mix left over for a full reaction. To avoid cross contamination, skip a position in each row. For example,
   Put 59 µL Cy3 UP mix in the first (no. 1) tube of each PCR strip.
   Put 59 µL Cy5 UP mix in the third (no. 3) tube of each PCR strip.
   Put 59 µL Cy3 DN mix in the fifth (no. 5) tube of each PCR strip.
   Put 59 µL Cy5 DN mix in the seventh (no. 7) tube of each PCR strip.
2. Take one 8-tube strip of PCR tubes from the "clean" area where you prepared the primer mixes and place it in a second prechilled rack located in a different area. Immediately cap the negative control strip. Otherwise, add 1 to 2 µL of Experimental (MM-URA) gDNA to the Cy3 UP (no. 1) and Cy3 DN (no. 5) tubes and 1 to 2 µL of the matched Control (MM+URA) gDNA to the Cy5 UP (no. 3) and Cy5 DN (no. 7) tubes (**Note 16**). Be sure to cap each row before getting another strip from the hood. Template gDNA concentrations must be *at least* 20 ng/µL, and preferably >200 ng/µL, to ensure representation of each TAG and to minimize sampling artifacts.
3. Place samples in a Perkin Elmer PE9600 (or comparable) PCR machine and initiate the following program:

Step	Temperature	Time
1	94°C	2 min
2	94°C	10 s
3	50°C	10 s
4	72°C	20 s
5	Return to **step 2** 49 times for a total of 50 cycles*	
6	4°C	

*If spurious hybridization to the *YQLnnnC* negative control features is observed, reduce the number of PCR cycles. In practice, as few as 35 cycles is usually sufficient.

4. A small amount (5 µL) of each TAG PCR reaction can be analyzed on a 3% agarose gel *(22)* or an 8% polyacrylamide gel and visualized using a fluorescence scanner or by EtBr staining. The expected PCR products should be ~54 to 58 bp; however, these products should be absent in a "No DNA" control PCR (**Note 17**).

3.7.5. Control Experiments

All of these suggested controls can be analyzed to identify problematic TAGs or features. They can reveal spurious background hybridization and also whether any Cy-labeled primers are contaminated by TAGs or otherwise showing differential behavior. In turn, this information can be exploited to create a filter applicable to all data sets generated using a given batch of primers. Use of these controls is further explained in Refs. *22, 29, 30*.

1. *"No DNA" template*. As noted above, each time PCR reactions are set up, one should include a "No DNA" Template Control. These mock extracts can be hybridized to a previously used microarray slide that has been stripped (see below). This control will reveal any spurious hybridization of fluorescent primers or primer dimers and/or whether any reagents are contaminated with TAG sequences. This is not a perfect control because it does not accurately mimic the experimental PCR conditions. A potentially better alternative that we have not yet explored is to use gDNA prepared from an isogenic wild-type *MAT***a** *can1Δ::LEU2-MFA1pr-HIS3* yeast strain that contains no UPTAGs or DNTAGs.
2. *Self-hybridization*. In addition, it is important to periodically perform a self-self hybridization experiment in which a single genomic DNA sample is used as template to make both the Cy3- and Cy5-labeled extracts. The gDNA sample can be chosen randomly. It is absolutely essential to do this control each time a new batch of primers is put to use. Hybridization of such samples will reveal whether any reagents (especially the Cy-labeled primers themselves) are contaminated or if the priming properties of the primer mixes are not equivalent. The aforementioned problems would be indicated if specific TAGs reproducibly show a higher signal intensity with one or the other Cy-labeled extract. Ideally, the normalized signal intensity ratio for each TAG should be ≈1.
3. *Dye-swap experiment*. It is also useful to periodically prepare and separately hybridize two sets of fluorescent extracts for a given experiment. The first set is prepared as usual, but for the second set, the fluorescent primers are "swapped" such that the Control (MM+URA) sample is Cy3-labeled and the Experimental (MM-URA) sample is Cy5-labeled.

3.8. Hybridization of Cy5- and Cy3-Labeled Fluorescent Extracts to Hopkins TAG Array or Other UPTAG and DNTAG Oligonucleotide Microarrays

Hybridization of fluorescent extracts to a microarray slide can be performed in either a heat-sealable bag (4 cm × 12 cm) or lidded compartment boxes, made of polypropylene (e.g., Alpha RHO Inc., cat. no. 776C-6-P). Most commercially available boxes are made of polystyrene, which strongly adsorbs DNA. For high-throughput dSLAM, boxes with five or six compartments simplify handling of the microarrays. Boxes should be rinsed several times with deionized water and dried prior to use. Note that both the hybridization buffer and the wash buffers are supplemented with DTT in an effort to counteract the detrimental effects of atmospheric ozone on Cy5.

3.8.1. Prehybridization

1. Wearing gloves, take out an Agilent microarray for each experiment (including the "No DNA" control) and record slide numbers on a data sheet. Place each slide in its own compartment of a multicompartment box. Be sure that the side labeled with the bar code plus the word *Agilent* is facing up. This is the side on which the oligonucleotide probes are arrayed (**Note 18**). Do not touch this surface.
2. Add 5 mL hybridization buffer (Hyb Bfr) freshly supplemented with 1 mM DTT to each compartment that contains a microarray. Avoid dispensing hybridization buffer directly onto the microarray surface.
3. In addition, for each microarray, aliquot 5 mL Hyb Bfr freshly supplemented with 1 mM DTT to a 15-mL conical tube.

4. Incubate both the boxes and the 15-mL tubes at 42°C for 30 to 60 min. The boxes should be gently rocked, but in a way that ensures the microarray slide surface is submerged at all times. This constitutes the *prehybridization step* and is optional. For production dSLAM, we routinely include the prehybridization step.

3.8.2. Hybridization

While the slides are prehybridizing, the fluorescent extracts that were prepared by PCR should be processed for hybridization. Throughout these steps, care should be taken to minimize ozone exposure and to shield the fluorescent materials from bright light.

1. For each set of PCR reactions, aliquot 15 μL of UPTAG Blocking Mix to one microfuge tube and 15 μL of DNTAG Blocking Mix to a separate microfuge tube.
2. Briefly centrifuge the PCR tubes or strips to collect liquid in the bottoms of the tubes. Next, for a given set of four PCR reactions, combine the Cy3 UP and Cy5 UP PCR samples together in a single microfuge tube that contains UPTAG Blocking Mix. Likewise, combine the Cy3 DN and Cy5 DN samples together in a single microfuge tube that contains DNTAG Blocking Mix (**Note 19**).
3. Vortex all tubes briefly and centrifuge to collect the liquid.
4. Incubate all UP and DN samples at 100°C for 2 min to denature double-stranded material, then transfer the tubes directly to ice. Cover ice bucket with a lid or aluminum foil to shield samples from light. Incubate samples on ice for at least 2 min before proceeding.
5. Add the combined and blocked UP and DN PCR reactions for a given experiment to one prewarmed 5-mL aliquot of Hyb Bfr. Mix well by gentle inversion. *Do not vortex*.
6. Retrieve boxes with slides from the 42°C incubator. Working one box at a time, aspirate the prehybridization solution. Do not touch the slide surface with the pipette. Pour each hybridization mixture into the compartment with the corresponding microarray slide. Remove or pop any large bubbles, as these can cause hybridization artifacts.
7. Wrap hybridization boxes in aluminum foil to prevent evaporation and to shield extracts from light. Incubate boxes with gentle rocking at 42°C overnight (at least 16 h).

3.8.3. Posthybridization: Washing and Scanning Microarray Slides

Thus far, we generally process hybridized microarray slides one at a time. This is to avoid certain artifacts that arise with batch processing (e.g., prolonged exposure to atmospheric ozone) ***(23)***. To wash, slides are sequentially immersed in wash buffers I and II. The slides are then spun dry in a minicentrifuge equipped with a slide adaptor (i.e., a slide spinner) and scanned.

1. For every two microarrays, prepare one 50-mL conical tube containing 50 mL wash buffer I: 6× SSPE supplemented with 0.05% Triton X-100 and 1 mM DTT (*freshly added*). For every single microarray, prepare one 50-mL conical tube containing wash buffer II: 0.06× SSPE freshly supplemented with 1 mM DTT.
2. Using Teflon-coated forceps or a wooden stick as a tool, remove one microarray slide from its hybridization compartment, and immediately transfer it to a 50-mL conical tube containing wash buffer I.
3. Using the forceps, draw the slide out of the wash buffer then gently drop it back into the tube. Repeat this “dunking” step three to five times.
4. Transfer the slide to a tube containing wash buffer II. Repeat the “dunking” step several times until the buffer begins to “sheet off” of the slide. At this point, the slide can be placed

in a slide spinner and spun dry for 10 s. Otherwise, the slide must be *very slowly* pulled out of the buffer such that no buffer clings to the slide.

5. The slide is then immediately placed into the slide chamber of a microarray scanner and data is collected. For the GenePix 4000B Scanner (Molecular Devices, Corp., Sunnyvale, CA) that we use, the labeled end of the slide should be toward the user and the side of the slide labeled with the word *Agilent* should be facing down (i.e., toward the laser and detector) (**Note 20**).

3.8.4. Microarray Data Acquisition

The following guidelines can be used to acquire data:

1. First, the laser power settings and photomultiplier tube (PMT) voltages should be set to maximize signal intensities without saturating any relevant features. With the protocols described here, a good starting point is 33% laser power and 600 volts PMT in both the F635 (Cy5) and the F532 (Cy3) channels. Excessive laser power can potentially bleach the fluorescence, although this has not been a problem in practice. Excessive PMT voltages will cause a large increase in background noise and must be avoided. Ideally, the settings chosen should produce unnormalized Cy5/Cy3 ratios that are close to 1 for most (but hopefully not all) features. However, this is not absolutely necessary, as the data can be normalized later.
2. Scanner resolution should be set to 10 μm/pixel. Scanning at a higher resolution offers no advantage. For each scan, we recommend saving the result as a multicolor image file (.tif).

3.9. Stripping Microarrays After Hybridization

A Hopkins TAG Array slide can be reused three to five times if carefully stripped and stored. This has obvious value, although usage should be tracked since stripping usually leaves several features (out of the 20,000+ on the slide) with discernible signal. For best results, a slide should be stripped immediately after it is scanned or at least on the same day as scanning.

3.9.1. Remove Adhesive Labels and Inscribe Bar Code

Working with one slide at a time, peel off the commercial bar-code labels from the slide. For bookkeeping purposes, it is useful to transfer the labels to your experimental data sheet. It is important to remove the labels because if the label ink leaches off during stripping, it can lead to a high green fluorescence background. After removing the label, use a diamond-point pen to inscribe the serial number on the slide. We routinely do this on the oligonucleotide array side of the slide. Finally, blow any visible dust (such as ink and glass debris from inscription) off the array with clean compressed air (e.g., “Dust-off ”).

3.9.2. Strip Slides

Place one or two such slide(s) into a 50-mL conical tube containing 50 mL 1% SDS, 10 mM EDTA, prewarmed to 75°C. Verify temperature with a clean thermometer. If stripping two slides in the same tube, be sure to place them back-to-back. Incubate the slide(s) in SDS/EDTA for 30-40 min at 75°C, with occasional agitation. Change the SDS/EDTA solution once during this incubation.

3.9.3. Rinse and Store Stripped Slides

After high-temperature incubation, rinse the stripped slides in wash buffers I and II as described above. At this stage, it is not necessary to supplement the buffers with DTT. To determine whether the stripping protocol was effective, re-scan each slide at low resolution (fast-scan). A second round of stripping is advised if more than several features retain discernible signal. Be sure to keep a record of the status of each slide. Spin-dry or air-dry the clean stripped slide, then store in a desiccator at 4°C or in a sealed container at −20°C for future use.

3.10. dSLAM Data Analysis and Follow-Up

The basic goal of dSLAM is to identify target genes that appear to have a synthetic lethal interaction with a given query gene. TAGs for which the normalized Cy5/Cy3 ratio is exceptionally high, meaning the TAGs are underrepresented in the double-mutant haploid pool, define potential target genes (**Fig. 3**). Methods for the analysis of TAG microarray data have been described elsewhere and range from a very simple Excel spreadsheet-based approach ***(15, 16)*** to ones that employ filtering and higher-level statistical manipulations (***[22, 29, 30]***; *see* **Chapter 25**).

Nonetheless, there are several controls that should be monitored to assess the efficacy of haploid selection on the two types of Magic Marker media (MM±URA) and the fidelity of hybridization. First, TAGs for any deletion allele that confers uracil auxotropy (e.g., YKOs for *URA1*, *URA2*, and *URA5* among others) should be underrepresented in the Experimental MM-URA pool because Ura$^-$ haploid strains should not grow in the absence of uracil. These TAGs routinely show high Cy5/Cy3 ratios, similar to a *bona fide* synthetic lethal target. Conversely, uracil transport defective mutants (e.g., the *FUR4/YBR021W* YKO haploid) do not grow well on MM+URA medium unless they are uracil prototrophs. In dSLAM experiments, these strains are relatively overrepresented in the MM-URA pool for which the *URA3MX*-marked query gene disruption allele was also selected, giving rise to a Cy5/Cy3 ratio much greater than 1. For reasons we do not fully understand, Cy5/Cy3 ratios much greater than 1 are also routinely observed for TAGs corresponding with the carbamoyl phosphate synthetase genes *CPA1/YOR303W* and *CPA2/YJR109C*, as well as the *CPA1*-upstream regulatory ORF *YOR302W*. Second, UPTAG and DNTAG signal intensities for the vast majority of essential genes should be relatively low in both haploid pools compared, for example, with that observed for the starting diploid pool. Third, YKO alleles that are linked to the query gene sometimes behave like synthetic lethal targets. This occurs when a query-adjacent deletion allele removes sequences required for efficient homologous recombination between the query gene disruption fragment and the chromosome. In this case, only the *YFG1* wild-type allele on the other homologue can be disrupted; thus, the probability of getting an *xxxN*Δ*::kanMX4 yfg1*Δ*::URA3MX* double mutant is directly related to the frequency of meiotic crossing-over between the query and query-adjacent genes. Fourth, UPTAG and DNTAGs corresponding with the query gene used for the experiment in question (i.e., *yfg1*Δ*::URA3MX*) should exhibit a significantly higher signal intensity than is observed in other experiments. This is because, in principle, ~50% of cells in the MM+URA pool and 100% of cells in the MM-URA pool now carry the query gene TAGs. Finally, signal intensities should be negligible for all negative control *YQLnnnC* features, as well as for features corresponding to strains absent from the starting pool or to TAGs that are severely broken (*22, 24, 30*; **Note 20**).

Once a candidate list of potential targets is obtained, it is up to the researcher to validate his or her results by attempting to construct the indicated double mutants through genetic crosses or by transformation. A method for pursuing the latter approach, 96 samples at a time, can be found in Refs. *15* and *16*. Obviously for high-throughput dSLAM, it will not be possible in the short-term to rigorously and directly validate each and every potential synthetic lethal combination. Thus, in the future we will rely on across-array analysis of ratio variance, on bidirectional analysis (i.e., X as query finds target Y; Y as query finds target X), and other bioinformatics data to confidently formulate candidate lists of synthetic lethal pairs. It will then be up to the greater yeast community-at-large to help us validate these lists. In turn, higher-order analysis of synthetic lethal data can be used to define epistasis groups, compensatory pathways, and cellular "wiring" maps *(15, 18, 20, 31, 32, 33, 34)*.

Note Added in Proof

Before this volume went to press, Agilent Technologies ceased manufacture of its single 22K arrays. Thus, the Hopkins TAG Array described above is no longer readily available. In response to the change in service, we designed a new higher-density multiplex version of the Hopkins TAG Arrays in collaboration with NimbleGen Systems, Inc. (Madison, WI). Each subarray of the 70K × 4 Array format (NimbleGen Design File 070323_DYuan.ndf; serial no. 5435) contains 5-fold replicates for most of the original UPTAG and DNTAG features, as well as 3-fold replicates for any remaining original features, for TAGs associated with new deletion strains (*11–13*) and for "corrected" sequences corresponding to broken TAGs (*24*). In addition, there are more negative control sequences (445 for each TAG type, up from 159). The multiplex format of the new design requires that some type of four-chambered, ultra-low volume, adhesive gasket be applied to the microarray slide to isolate each subarray. Currently, we are optimizing a new nltra-low volume hybridization protocol and evaluating results obtained using NimbleChip X4 Mixers (NimbleGen Systems, Inc. cat. no. 0038G4), in conjunction with the MAUI Hybridization System (BioMicro Systems, Inc., Salt Lake City, Utah). Once finalized, our new hybridization protocol will be available at the dSLAM project web site (http://slam.bs.jhmi.edu/). Likewise, once published, a revised hoptag software package (version 3) capable of performing a full analysis on datasets obtained with either the Agilent or NimbleGen platform will be freely available at http://slam.bs.jhmi.edu/hoptag/ under a standard license.

Notes

1. We routinely use *URA3* as the query gene disruption selectable marker, in part because *URA3* is the only prototrophic marker available for selection in the Magic Marker heterozygous deletion diploid background. In addition, the stability of query gene disruption transformants can be assessed by replica printing the primary Ura$^+$ transformants to plates containing 5-fluoroorotic acid (**Section 3.4.4**, **step 6**). However, other dominant drug-resistance markers can also be used, although obviously the media requirements will change

accordingly. For example, the *NatMX4* cassette allows for expression of nourseothricin *N*-acetyltransferase in budding yeast and confers resistance to the antibiotic ClonNAT (*21*).

For the purposes of high-throughput dSLAM, we derived a collection of *MAT***a** YKO haploid strains in which the *kanr* sequence of *kanMX4* cassette has been replaced with the *URA3* gene by homologous recombination. The "Magic Marker" *yfg*$_N$Δ*::URA3MX/YFG*$_N^+$ heterozygous deletion diploid progenitors of these *MAT***a** *yfg*$_N$Δ*::URA3MX* strains are now available from Open Biosystems (cat. no. YSC4428). The detailed construction of this strain collection will be described later. Individual query gene disruption fragments are obtained by PCR amplification of the *yfg*$_N$Δ*::URA3MX* cassette plus ~1.5 kb of 5′ and 3′ flanking DNA genomic using gene-specific primers and genomic DNA prepared from the corresponding viable haploid as a template. As the desired PCR products are ~5 kb in length, we use LA Taq from Takara Bio USA. This DNA polymerase has 3′ to 5′ exonuclease (proofreading) activity and is especially effective in amplifying longer PCR fragments. However, a query gene disruption fragment with at least 1.5 kb of flanking homology can also be generated by conventional cloning techniques.

2. Refer to **Figure 1** for more information regarding oligonucleotides. As discussed in **Section 3.6.2** and also in Refs. *16* and *22*, extreme care should be taken to avoid contaminating dSLAM PCR reagents with spurious TAG sequences. Ideally, all oligonucleotides should be processed in a "clean room" or "clean hood" setting, using dedicated equipment that has never been exposed to yeast or yeast DNA. Also, cyanine dyes—especially Cy5—are subject to signal degradation by atmospheric ozone (*23*). Thus, it is desirable to work with these fluorophores (as well as the microarrays to which they have been hybridized) in a controlled environment. Monitor ozone levels in your lab. If ozone levels routinely exceed 10 ppb, then you should install an air-filtration system to remove the ozone.
3. Hopkins TAG Arrays have 22,575 features, including probes for all UPTAGs and DNTAGs present in YKO strains that comprise the heterozygous deletion diploid collection mentioned above. In addition, UPTAGs and DNTAGs for 800 genes are replicated an additional five times, as are UPTAGs and DNTAGs for 159 nonexistent ORFs that serve as independent negative controls (annotated as "YQLnnnC" in array documentation). Academic researchers who are interested in using Hopkins TAG Arrays should first join the Yeast Barcode Array Consortium (http://barcode.princeton.edu/) in order to receive an academic discount. In the future, we plan to redesign the Hopkins TAG Array to include UPTAG and DNTAG probes for newly discovered genes and corrected UPTAG and DNTAG sequences for strains in which the actual TAG sequence differs from the intended TAG sequence (so-called broken tags) (*24*). Refer to the Yeast Barcode Array Consortium Web site or our public dSLAM Web site (http://slam.bs. jhmi.edu/) for updates in this regard. The Hopkins TAG Array design is available from the Gene Expression Omnibus (GEO) microarray data repository at NCBI (accession no. GPL1444). Annotation files for use with various scanner platforms (GMEL, MAGE-ML, and GAL) can be obtained through either the consortium or dSLAM Web site. Microarrays are provided from Agilent in a slide box in a vacuum-sealed foil pouch. Sealed packages can be stored at room temperature for several months. However, we have observed erratic performance with slides from opened packages. Therefore, we recommend storing any unused slides in either a desiccator at 4°C or a sealed bag at −20°C.
4. Wash buffers can be prepared in advance and aliquots made to 50-mL conical tubes, except that DTT should be omitted and added just before use.
5. Prepare at least two copies of each plate in case contamination occurs. At this time, if desired, the strains can also be pinned onto various media to confirm diploid phenotypes are as expected (i.e., nonmating, Ura$^-$ Leu$^-$ His$^-$ Can^S).

6. Based on mathematical considerations as well as empirical tests, the experiment should be discarded if a minimal number of 3×10^5 transformants is not obtained. Otherwise random "dropouts" can occur and representation of the 6000 heterozygous deletion strains present in the initial pool degrades.
7. *Assess the stability of query gene disruption transformants*. For each transformation, replica print one of the SC-URA-LEU plates used to titer transformation efficiency to an SC-LEU+5-fluoroorotic acid (5-FOA) plate. Patch stable *URA3*$^+$ and *ura3*Δ strains on the 5-FOA plates as controls. 5-FOA is a toxic pyrimidine analogue that selects against cells that are wild type for *URA3* (**25**). Ura$^+$ colonies that also grow well in the presence of 5-FOA are unstable transformants, perhaps owing to the presence of origin or ARS (autonomously replicating sequence) activity on the query gene fragment. Thus, it is unlikely that the genomic copy of the query gene has been disrupted in these colonies. In this case, only proceed if the number of *stable* (i.e., 5-FOA^S) transformants is still $\geq 3 \times 10^5$. If not, the query gene disruption fragment should be redesigned to exclude probable ARS activity (**26, 27**). If the experiment is carried forward, it is important to test the stability of the Ura$^+$ haploids selected after sporulation (**Section 3.5**). If a significant number of Ura$^+$ haploids are also unstable, the experiment should be aborted because the putative double-mutant population will contain single xxx_N*Δ::kanMX4* mutants that are Ura$^+$ by virtue of an unstable derivative of the *yfg1Δ::URA3MX* fragment.
8. *Utility of storing multiple aliquots of double heterozygous deletion diploids.* Such pools of double heterozygous deletion diploids are a valuable resource for several reasons. First and most obviously, they serve as a backup if sporulation or haploid selection fails for some reason (e.g., contamination occurs). One can simply thaw a reserve aliquot of diploid transformants and repeat the sporulation and selection protocol. Likewise, one might wish to change the parameters of haploid selection (e.g., vary the temperature or add or omit a nutrient or drug from the media). Also, one might want to compare the representation of strains in a transformed diploid pool to that in the parental Magic Marker xxx_N*Δ::kanMX4/*XXX_N^+ pool or in the derived haploid pools. The archived aliquots are a perfect source of cells for genomic DNA isolation. Finally, one might want to search for triple mutant synthetic lethality, in which case a particular double heterozygous deletion diploid pool can be transformed with a second query construct (e.g., *yfg2Δ::NatMX*).
9. The total ODUs per culture is the $OD_{600} \times 2$ (dilution factor) × 50. Simply divide the total ODUs by 10 or multiply the OD_{600} of the 1 : 1 dilution by 10 to calculate the appropriate volume.
10. Ideally, the number of Ura$^+$ haploid colonies on MM-URA medium should be approximately one-half the total number of *MAT***a** xxx_N*Δ::kanMX4* colonies obtained on MM+URA medium (or equal to the number of Ura- colonies that grew on MM+URA+5-FOA). However, the proportion of Ura$^+$ colonies can be less than expected if, for example, there is nonrandom spore inviability associated with the *yfg1Δ::URA3MX* query gene disruption.
11. Storage of haploid pools in water (or as aspirated cell pellets) is ideal for subsequent genomic DNA preparations (**Section 3.6**). However, if a need for viable haploid pools is anticipated, then glycerol should be added to the cell suspension to a final concentration of ~15% prior to freezing.
12. *Recommended precautions for gDNA preparation and subsequent PCR reactions*. We have found that extreme care must be taken to avoid "environmental" and cross-contamination of samples during gDNA preparation and steps relevant to the PCR amplification that follows (e.g., dilution and aliquoting of primers, setting up reactions, etc.). Such contamination is manifest as unexpected hybridization signals. For this reason, we routinely prepare

gDNAs for dSLAM in a "clean room" setting; that is, one in which yeast knock-out strains have never been handled except for dSLAM-related gDNA or PCR preparative steps and for which "virgin" equipment has been purchased, most especially the racks, pipettors, and microcentrifuges. Filter pipet tips are used at all times and gloves are changed frequently. Moreover, we highly recommend that pipette tips be changed for each and every pipetting that involves tubes that contain cells or gDNA—even when aliquoting the same general reagent into a series of tubes. For additional information, the reader is directed to the Supplementary Material in Ref. *22*.

13. Thoroughly mix the Yeast Cell Lysis Solution to ensure uniform composition before dispensing. Also, we do not use the RNase A provided in the Epicentre kit. Rather we use 20 mg/mL RNase A purchased from Invitrogen.
14. Refer to the QIAamp DNA Micro handbook P41 or www.qiagen.com for additional information. Do not use the carrier RNA for dSLAM!
15. For long-term storage, EDTA can be added to 1 mM final concentration. However, the presence of EDTA might interfere with subsequent UPTAG and DNTAG PCR amplification.
16. Because the Cy5 fluorophore signal is susceptible to ozone degradation, it is recommended that Cy5-labeled primers routinely be used to amplify Control samples, the exception being purposeful dye-swap experiments.
17. Ideally, no PCR product should be observed in the "No DNA" template controls. A 30-bp product is commonly seen instead, attributable to primer dimers formed by either the U2c or D2c primer, each of which have a six-base restriction site at the 3′ end *(22)*. Contamination of reagents by individual UP or DN tags will lead to production of ~54- to 58-bp PCR products in the "No DNA" template controls.
18. To avoid cross-contamination of extracts, hybridizations should be performed only in every other compartment of the multichambered box (i.e., three microarray slides per box).
19. The blocking oligonucleotides are designed to anneal to the universal priming sites on the single-stranded Cy-labeled PCR products. In this way, only the TAG-complementary sequences remain single-stranded and available to hybridize to the TAG sequences displayed on the microarray. In principle, this step should reduce spurious hybridization.
20. *Missing strains and broken tags.* A substantial fraction of TAGs (roughly 35%) will consistently have signal intensities at or near background levels. There are several explanations for this, including (a) absence of certain heterozygous YKO diploid strains from the starting pool; (b) failure of certain YKO diploid strains to convert to the haploid state; (c) inability to select *MAT***a** haploids on Magic Marker media owing to defects in mating type identity, histine or leucine auxotrophy, or failure of the Magic Marker; (d) pronounced growth defects associated with certain YKO alleles that lead to TAG underrepresentation in freshly converted YKO haploid pools (e.g., ~1100 haploids deleted for essential genes); (e) mutations in the universal priming sequences or the TAGs themselves that prevent their PCR amplification or hybridization to the array *(24)*. Presumably, class A and some instances of classes C and D can be predicted ahead of time. Classes A and E can be distinguished from the others by hybridizing TAGs amplified from gDNA corresponding with the starting heterozygous YKO diploid pool.

Acknowledgments

This work was supported in part by NHGRI grant HG02432 and by NIH Roadmap grant RR020839 to J.D.B. X.P. was a fellow of the Leukemia and Lymphoma Society.

D.Y. was also supported in part by a fellowship from the Burroughs-Wellcome Center for Computational Biology at Johns Hopkins University and by the Ruth L. Kirschstein NRSA fellowship from NHGRI. The authors thank Siew-Loon Ooi for her early contributions to the development of the SLAM technology and Yu-Yi Lin, Emily Kleinhans, and Abid Khan for their ongoing contributions.

Under a licensing agreement between Open Biosystems, Inc. and the Johns Hopkins University, the University is entitled to a share of royalties on sales of yeast strains described in this paper. The terms of this arrangement are being managed by the Johns Hopkins University in accordance with its conflict of interest policies.

References

1. Brownstein M.J., Khodursky, A., and Conniffe, D. B. (2003) *Functional Genomics: Methods and Protocols. Methods in Molecular Biology.* Totowa, NJ: Humana Press; **224**.
2. Pevsner, J. (2003) *Bioinformatics and Functional Genomics*. Hoboken, NJ: John Wiley & Sons, Inc.
3. Oliver, S. G., van der Aart, Q. J., Agostoni-Carbone, M. L., Aigle, M., Alberghina, L., Alexandraki, D., et al. (1992) The complete DNA sequence of yeast chromosome III. *Nature* **357**, 38–46.
4. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldman, H., et al. (1996) Life with 6000 genes. *Science* **274**, 546–567.
5. Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., et al. (1999) Functional characterization of the *Saccharomyces cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906.
6. Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., et al. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome, *Nature* **418**, 387–391.
7. Scherens, B., and Goffeau, A. (2004) The uses of genome-wide yeast mutant collections. *Genome Biol.* **5**, 229.
8. Wach, A., Brachat, A., Poehlmann, R., and Philippsen, P. (1994) New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*. *Yeast* **10**, 1793–1808.
9. Ooi, S.-L., Shoemaker, D. D., and Boeke, J. D. (2001) A DNA microarray-based genetic screen for nonhomologous end-joining mutants in *Saccharomyces cerevisiae*. *Science* **294**, 2552–2556.
10. Pan, X., Yuan, D. S., Xiang, D., Wang, X., Sookhai-Mahadeo, S., Bader, J. S., et al. (2004) A robust toolkit for functional profiling of the yeast genome. *Mol. Cell* **16**, 487–496.
11. Kessler, M. M., Zeng, Q., Hogan, S., Cook, R., Morales, A. J., and Cottarel, G. (2003) Systematic discovery of new genes in the *Saccharomyces cerevisiae* genome. *Genome Res.* **13**, 264–271.
12. Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254.
13. Kastenmayer, J. P., Ni, L., Chu, A., Kitchen, L. E., Au, W. C., Yang, H., et al. (2006) Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae. Genome Res.* **16**, 365–373.
14. Ooi, S.-L., Pan, X., Peyser, B. D., Ye, P., Meluh, P. B., Yuan, D. S., et al. (2006) Global synthetic-lethality analysis and yeast functional profiling. *Trends Genet.* **22**, 56–63.
15. Pan, X., Ye, P., Yuan, D. S., Wang, X., Bader, J. S., and Boeke, J. D. (2006) A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell* **124**, 1069–1081.

16. Pan, X., Yuan, D. S., Ooi S.-L., Wang, X., Sookhai-Mahadeo, S., Meluh, P. and Boeke, J. D. (2007) dSLAM analysis of genome-wide genetic interactions in *Saccharomyces cerevisiae*. *Methods* **41**, 206–221.
17. Tong, A. H., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Page, N., et al. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368.
18. Tong, A. H., Lesage, G., Bader, G. D., Ding H., Xu, H., Xin, X., et al. (2004) Global mapping of the yeast genetic interaction network. *Science* **303**, 808–813.
19. Tong, A. H., and Boone, C. (2006) Synthetic genetic array analysis in *Saccharomyces cerevisiae*. *Methods Mol. Biol.* **313**, 171–192.
20. Schuldiner, M., Collins, S. R., Thompson, N. J., Denic, V., Bhamidipati, A., Punna, T., et al. (2005) Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123**, 507–519.
21. Goldstein, A. L., and McCusker, J. H. (1999) Three new dominant drug resistance cassettes for gene disruption in *Saccharomyces cerevisiae*. *Yeast* **15**, 1541–1553.
22. Yuan, D. S., Pan, X., Ooi, S.-L., Peyser, B. D., Spencer, F. A., Irizarry, R. A. and Boeke, J. D. (2005) Improved microarray methods for profiling the Yeast Knockout strain collection. *Nucleic Acids Res.* **33**, e103.
23. Fare, T. L., Coffey, E. M., Dai, H., He, Y. D., Kessler, D. A., Kilian, K. A., et al. (2003) Effects of atmospheric ozone on microarray data quality. *Anal. Chem.* **75**, 4672–4675.
24. Eason, R. G., Pourmand, N., Tongprasit, W., Herman, Z. S., Anthony, K., Jejelowo, O., et al. (2004). Characterization of synthetic DNA bar codes in *Saccharomyces cerevisiae* gene-deletion strains. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11046–11051.
25. Boeke, J. D., Trueheart, J., Natsoulis, G., and Fink, G. R. (1987) 5-Fluoroorotic acid as a selective agent in yeast molecular genetics. *Methods Enzymol.* **154**, 164–175.
26. Breier, A. M., Chatterji, S., and Cozzarelli, N. R. (2004) Prediction of *Saccharomyces cerevisiae* replication origins. *Genome Biol.* **5**, R22.
27. Feng, W., Collingwood, D., Boeck, M. E., Fox, L. A., Alvino, G. M., Fangman, W. L., et al. (2006) Genomic mapping of single-stranded DNA in hydroxyurea-challenged yeasts identifies origins of replication. *Nat. Cell Biol.* **8**, 148–155.
28. Hoffman, C. S., and Winston, F. (1987) A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of *Escherichia coli. Gene* **57**, 267–272.
29. Peyser, B. D., Irizarry, R. A., Tiffany, C. W., Chen, O., Yuan, D. S., Boeke, J. D., and Spencer F. A. (2005) Improved statistical analysis of budding yeast TAG microarrays revealed by defined spike-in pools. *Nucleic Acids Res.* **33**, e140.
30. Peyser, B. D., Irizarry, R., and Spencer, F. A. (2006). Statistical analysis of fitness data determined by TAG hybridization on microarrays. This volume.
31. Ye, P., Peyser, B. D., Spencer, F. A., and Bader J. S. (2005) Commensurate distances and similar motifs in genetic congruence and protein interaction networks in yeast. *BMC Bioinformatics* **6**, 270.
33. Ye, P., Peyser, B. D., Pan, X., Boeke, J. D., Spencer, F. A., and Bader J. S. (2005) Gene function prediction from congruent synthetic lethal interactions in yeast. *Mol. Syst. Biol.* **1**, 2005.0026.
34. Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J., Hon, G. C., Myers, C. L., et al. (2006) Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae. J Biol.* **5**, 11.

16

Scarless Engineering of the *Escherichia coli* Genome

Tamás Fehér, Ildikó Karcagi, Zsuzsa Győrfy, Kinga Umenhoffer, Bálint Csörgő, and György Pósfai

Summary

E. coli K-12, being one of the best understood and thoroughly analyzed organisms, is the workhorse of genetic, biochemical, and systems biology research, as well as the platform of choice for numerous biotechnological applications. Genome minimization/remodeling is now a feasible approach to further enhance its beneficial characteristics for practical applications. Two genome engineering techniques, a lambda Red–mediated deletion method and a suicide (conditionally replicative) plasmid-based allele replacement procedure are presented here. These techniques utilize homologous recombination, and allow the rapid introduction of virtually any modifications in the genome.

Key Words: double-strand break; *Escherichia coli*; lambda Red recombinase; recombination; I-SceI endonuclease; serial genome modification; suicide plasmid.

1. Introduction

In the postgenomic era, large-scale remodeling of bacterial genomes became possible. The search for a minimal genome, metabolic engineering, and design of cell factories are projects that require multiple, serial modifications of the genome. Ideally, the genome-engineering techniques used for such tasks should be simple, relatively high-throughput, and leave no exogenous DNA segments behind. Littering the genome with remnants of the constructs (e.g., recombinase target sites, marker genes) can result in polar effects or genomic rearrangements and can complicate further manipulations.

Based on comparative genomics, about 80% of the genes of *Escherichia coli* K-12 MG1655 were identified as the "core" *E. coli* genome present in all strains for which the genome sequence is available ***(1)***. The remaining 20%, corresponding with about 100 segments, represent strain-specific genomic islands, loaded with prophages, insertion elements, and genes with unknown functions. With the ultimate goal to construct an *E. coli* strain with the core genome, we achieved a substantial (>15%) reduction/correction of the *E. coli* K-12 genome ***(2)***. Multiple deletions, insertions, and other modifications were sequentially introduced in the genome, using homologous

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

recombination-based DNA-modifying techniques. We describe here two genome-engineering methods: (a) modified, lambda Red–mediated, linear DNA–based deletion technique *(3)* and (b) suicide (conditionally replicative) plasmid-based allele replacement *(4)*. Whereas the first method is used mainly for constructing serial deletions, the second method is preferred for small, precise genomic surgeries (insertions, point mutations, reconstruction of interrupted genes). Both methods employ a double-strand break (DSB)-stimulated resolution of intermediate DNA constructs, resulting in scarless (devoid of exogenous sequences) genomic modifications.

Both methods require the use of a recombination-proficient *E. coli* host. The linear DNA–based method, in addition, employs lambda Red recombinase functions expressed from a plasmid *(5)*. Acting on short, homologous target sequences, Red recombinases can be used for straightforward replacement of a genomic segment with a marked exogenous sequence, flanked by synthetically produced DNA segments *(5–7)*. The suicide plasmid–based method is more labor-intensive, requiring cloning of longer homologous targeting fragments in the plasmid vector, followed by integration of the plasmid in the chromosome.

In both cases, the second step of the procedure is a RecA-mediated resolution of the intermediate genomic constructs, stimulated by targeted cleavage of the chromosome by meganuclease I-SceI *(8)*. In the case of linear DNA–based method, the end product of the procedure should be the desired scarless deletion. The suicide plasmid–based method, however, can result either in wild-type (wt) or mutant genome, and thus a polymerase chain reaction (PCR) screen must be employed to select the desired clones.

E. coli K-12 is one of the best understood and thoroughly analyzed organisms. Currently, 87% of the genes have functional assignments *(9)*. However, because much of the K-12 protein interactome remains obscure, predicting the effects of gene deletions/modifications is not trivial, and unexpected results of multiple genomic modifications (e.g., synthetic lethals or synthetic beneficials) *(2, 10, 11)* are likely. In this respect, the two alternative methods presented here are not equally informative. Failure by the Red-mediated deletion method can indicate either lethality of the planned construct or a technical problem. Using the suicide plasmid–based method, more information can be obtained. "One-sided" resolution of the intermediate construct or unexpected ratio of wild type (wt) and mutant end products can indicate more subtle physiologic changes as well.

2. Materials

1. Recombination-proficient *E. coli* strain.
2. pSG76-CS plasmid *(3)*, (GenBank accession no. AF402780).
3. pST76-A/C/K *(12)*, (accession nos. Y09895, Y09896, Y09897, respectively) or similar suicide plasmid vector.
4. pKD46 plasmid *(5)* (AY048746).
5. pSTKST plasmid *(3)* (AF406953).
6. Standard microbial growth media. LB (Luria-Bertani) medium: 1% bacto-tryptone, 0.5% bacto-yeast extract, 1% NaCl pH7.0.

   SOC medium: 2% bacto-tryptone, 0.5% bacto-yeast extract, 0.05% NaCl pH7.0. After sterilization, add sterile glucose to achieve 20 mM, and sterile $MgCl_2$ to obtain 10 mM final concentrations *(13)*.

7. Antibiotics ampicillin (Ap), chloramphenicol (Cm), and kanamycin (Km), used at a final concentration of 100 µg/mL, 50 µg/mL, and 50 µg/mL, respectively.
8. Chlortetracycline-hydrochloride (CTc): Add 30 mg CTc to 30 mL LB. Sterilize by autoclaving for 20 min at 15 lb/sq. in. on liquid cycle. Store at 4°C, protected from light. Use at a final concentration of 30 µg/mL.
9. L-(+)-Arabinose: Dissolve 1 g L-(+)-arabinose in 10 mL deionized H_2O and filter across a sterile Millex GP filter with 0.22-µm pore size. Use at a final concentration of 0.1%.
10. TE (Tris-EDTA buffer): Dissolve 1.21 g Tris(hydroxymethyl)aminomethane and 0.292 g EDTA in 1.000 mL H_2O. Adjust to pH 7.4 with HCl.
11. Restriction enzymes, T4 DNA ligase, *Taq* DNA polymerase, or other PCR enzymes, and buffers supplied by the manufacturers.
12. Agarose gel electrophoresis equipment.
13. Oligonucleotide primers.
14. PCR purification kit.
15. Bacterial electroporation equipment.

3. Methods

3.1. Lambda Red–Mediated, Linear DNA–Based Deletion Method

Steps of the method are depicted in **Figure 1**. To delete the chromosomal region between arbitrarily chosen segments (A, B), a composite linear DNA molecule is generated by PCR on plasmid pSG76-CS. Oligonucleotide primers (**ab**, **c**) with 5′-extensions provide the terminal "homology boxes" required for recombination into the genome

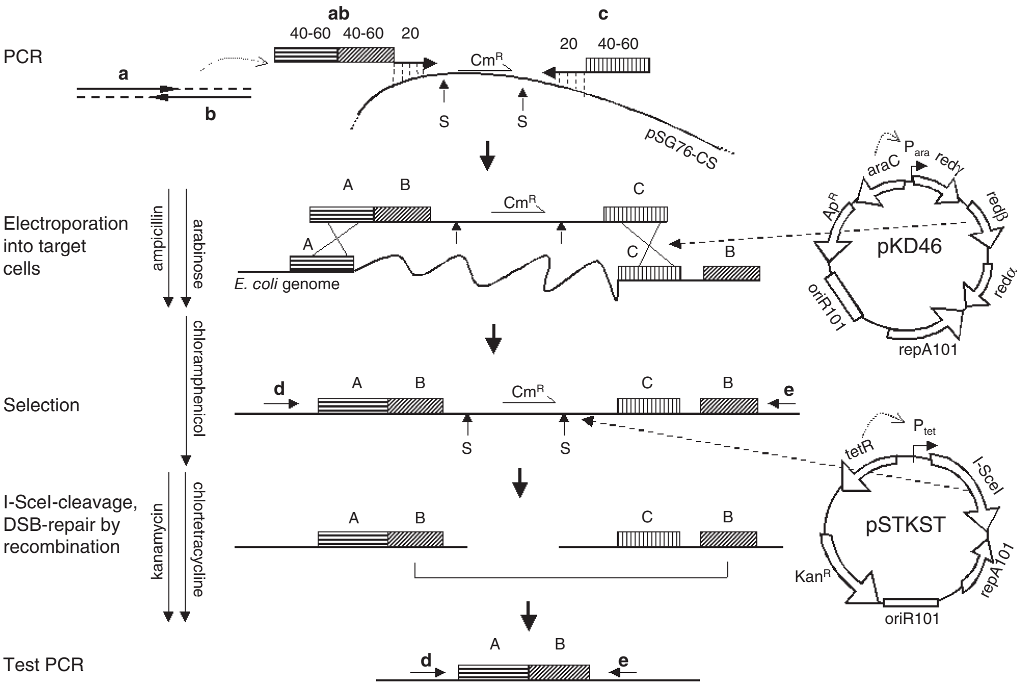


Fig. 1. Overview of the linear DNA–based deletion method. **A**, **B**, and **C** represent arbitrarily chosen DNA segments (homology boxes). Numbers refer to the length of the oligonucleotides. PCR primers are labeled by lowercase letters (**a**, **b**, **c**, **d**, **e**, **ab**). **S** indicates an I-SceI cleavage site.

(**Note 1**). (Because synthesis of very long primers is difficult, primer **ab** is generated in a PCR-like filling-in reaction of two partially complementary oligonucleotides.) The fragment, carrying a selectable marker (chloramphenicol resistance, Cm^R) flanked by I-SceI sites, is electroporated into the cell, where it can replace a segment of the chromosome via double crossover involving the terminal "homology boxes" A and C. The helper plasmid pKD46 provides the arabinose-inducible recombinase functions. Cells that integrated the linear fragment in the chromosome are selected by their Cm^R phenotype. Next, I-SceI expression is induced from helper plasmid pSTKST, resulting in cleavage of the chromosome at the 18-bp recognition sites present on the integrated fragment. As the broken ends carry short homologous regions (box B) close to their termini, RecA-mediated intramolecular recombinational repair preferentially proceeds via these segments (**Notes 2** and **3**). Generation of the desired, scarless deletion is confirmed by PCR using flanking primers (**d**, **e**).

3.1.1. Generation of the Targeting Fragment

1. Mix 20 pmol primer **a** with 20 pmol primer **b** and perform PCR in a total volume of 50 µL. Run 15 cycles with parameters 94°C 40 s/57°C or lower (depending on the complementarity of **a** and **b**) 40 s/72°C 15 s.
2. Mix 1 µL of the PCR product with 20 pmol each of primers **a** and **c** and 50 ng of pSG76-CS template and perform a second round of PCR. Run 28 cycles at 94°C 40 s/57°C 40 s/72°C 80 s.
3. Purify the PCR-generated fragment with a PCR purification kit and suspend it in 20 µL water. Elimination of the template plasmid (e.g., by *Dpn*I digestion) is not needed.

3.1.2. Insertion of the Targeting Fragment in the Genome

1. Grow the target *E. coli* strain harboring pKD46 at 30°C in 50 mL LB+Ap medium. Add 0.1% L-arabinose at early logarithmic phase and harvest the culture at OD_{600} of ~0.6. Prepare electrocompetent cells by concentrating the culture 100-fold, washing three times with ice-cold water, and resuspending in 10% glycerol (*13*).
2. Electroporate 4 µL targeting DNA fragment (100 to 500 ng) into 40 µL of electrocompetent cells.
3. Add shocked cells to 1 mL SOC, incubate 1 to 2 h at 37°C, then sediment cells by a brief spin in a microcentrifuge, spread them on LB+Cm plates, and incubate at 37°C. Expect 10 to several hundred colonies to appear in 36 h, 5% to 90% of which contain the desired insertion.
4. Check insertion of the fragment by colony PCR using primers **d** and **e**. (Touch the colony with a sterile toothpick, drop the toothpick in an Eppendorf tube containing 20 µL TE, vortex briefly, and use 1 µL of the cell suspension as PCR template.)

3.1.3. Removal of the Exogenous Sequences from the Genome

1. Transform cells harboring the desired insertion by pSTKST using standard transformation protocols (*13*). Spread cells on LB+Km plates and incubate at 30°C.
2. Inoculate a colony into 10 mL of LB+Km medium supplemented with heat-treated CTc (25 µg/mL final concentration) and grow for 24 to 36 h at 30°C.
3. Plate 10^{-5} to 10^{-6}-fold dilutions of the culture on LB+Km+CTc plates and incubate for 12 to 24 h at 30°C.

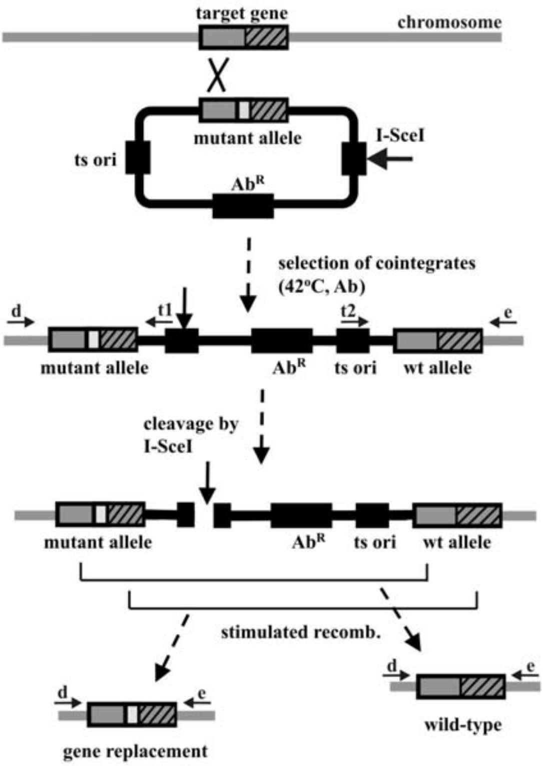


Fig. 2. General scheme of the suicide plasmid–based allele replacement procedure. A mutation between homology arms, cloned in a suicide plasmid, is indicated by an empty box. **Ab** stands for antibiotic resistance gene, **ts** represents a temperature-sensitive replicon, and an **arrow** indicates an I-SceI cleavage site. I-SceI is inducibly expressed from helper plasmid pSTKST (not shown). PCR primers are labeled by lowercase letters (**d**, **e**, **t1**, **t2**). (Adapted from Ref. *4* by permission of Oxford University Press.)

4. Screen for loss of insertion in 6 to 12 colonies by colony PCR using primers **d** and **e** (**Notes 4** and **5**).

3.2. Suicide Plasmid–Based Genome Modifications

The method is outlined in **Figure 2**. A suicide plasmid, carrying a targeting DNA fragment with the planned mutation (deletion, insertion, point mutation) in its middle (**Note 6**) is transformed in the cell. The plasmid can integrate into the chromosome via single crossover involving one of the “homology arms” of the mutant allele and the corresponding chromosomal region. Such cointegrates are selected by their antibiotic resistance at the nonpermissive temperature for plasmid replication. Next, I-SceI expression is induced from helper plasmid pSTKST, resulting in cleavage of the chromosome at the 18-bp recognition site present on the integrated plasmid. RecA-mediated intramolecular recombinational repair of the chromosomal gap utilizing the homologous segments close to the broken ends can result either in a reversion to the wt chromosome or in a markerless allele replacement.

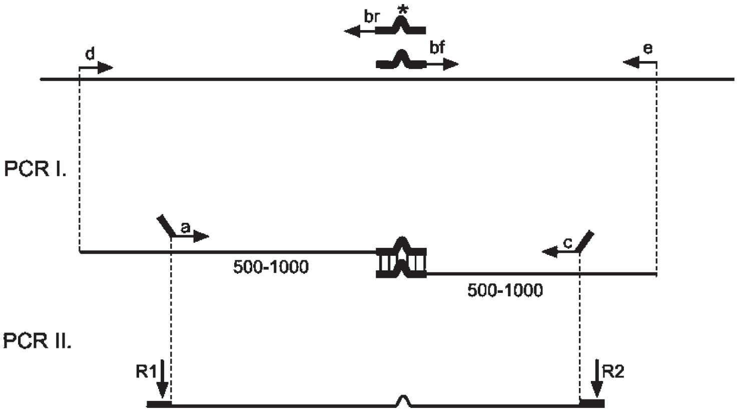


Fig. 3. A routine procedure for generating a targeting fragment for cloning in a suicide plasmid. Numbers refer to the lengths of nucleotide sequences. Arrows and lowercase letters indicate PCR primers. Primers **a** and **c** carry 5′-extensions with unique restriction enzyme recognition sites (**R1**, **R2**). Primers **br** and **bf** carry the mutation (deletion, insertion, point mutation) to be introduced in the chromosome in their complementary 5′ ends, marked by an asterisk (*). The two-step PCR procedure is described in the text.

3.2.1. Generation and Cloning of the Composite Targeting Fragment

Using standard recombinant PCR methods, generate a composite DNA fragment harboring the desired mutation in the middle of the fragment, flanked by arms with at least 500-bp homology with the targeted chromosomal region. Clone the fragment in a pST76-type *(12)* or similar suicide plasmid. Details of a routine procedure (**Fig. 3**), used frequently in our laboratory, are given below.

1. Mix 20 pmol primer **d** with primer **br** and perform PCR on 1 µg genomic DNA template in a 50-µL volume. Cycle parameters are typically 28× (94°C 40 s/57°C 40 s/72°C 60 s). Similarly, perform PCR using primers **bf** and **e**. Primers **br** and **bf** carry 5′ overhangs that are complementary to each other.
2. Purify both PCR products with a PCR fragment purification kit and suspend them in 20 µL water each. Combine 2 µL of each in a tube, add 20 pmol of primers **a** and **c** each, and perform PCR as above.
3. Purify the PCR product and cleave with restriction enzymes **R1** and **R2**.
4. Purify the fragment and ligate it in the appropriately cleaved multiple cloning site of pST76-A (*12*). Transform cells with the ligated mixture, spread the culture on selective (LB+Ap) agar plates, and incubate at 30°C.

3.2.2. Insertion of the Suicide Plasmid in the Chromosome

1. Transform the suicide plasmid carrying the mutant allele into the target cell, spread the culture on agar plates supplemented with the appropriate antibiotic, and incubate at 30°C for 12 to 24 h.
2. Pick four colonies and restreak them on fresh selective plates. Incubate the plates at 30°C for 2 to 6 h.
3. Transfer the plates to 42°C and incubate them for 12 h. Next, transfer the plates to 37°C and incubate them for an additional 12 to 24 h. Typically, some large colonies or sectors

of colonies are formed over the background of small colonies. These large colonies carry the plasmid integrated into the chromosome.

4. Pick a few large colonies, restreak them on selective agar plates, and incubate at 37°C for 12h. The arising colonies should be uniform in size.
5. Verify the site of insertion by colony PCR using appropriate primer pairs (**d**,**t1**; **t2**,**e**; *see* **Fig. 2**).

3.2.3. Removal of the Exogenous Sequences from the Genome and Verification

1. Transform the cells harboring the desired plasmid cointegrate with pSTKST, using standard transformation protocols *(13)*. Spread cells on LB+Km plates and incubate at 30°C.
2. Inoculate a colony into 10mL of LB+Km medium supplemented with heat-treated CTc (25 µg/mL final concentration) and grow for 24h at 30°C.
3. Plate 10^{-5} to 10^{-6}-fold dilutions of the culture on LB+Km+CTc plates and incubate for 12 to 24h at 30°C.
4. Screen typically 12 to 24 colonies by colony PCR using primers appropriate for distinguishing wt and mutant alleles. In the case of small mutations with no physiologic effect under the growth conditions applied, expect a 50:50 ratio of wt and mutant colonies (**Notes 4, 5,** and **7**).

Notes

1. For the linear DNA–based method, the terminal "homology boxes" should be 40 to 65bp long. Longer homologies usually allow more efficient integration into the chromosome; however, we observed large variations in recombination efficiency among constructs with similar lengths of homology. For reasons not completely understood, different regions of the chromosome seem to display very different recombinogenic potentials.
2. In the course of our work, deletions ranging from a few bp to 82kb were generated by the linear DNA–based method. The 1.7-kb linear fragment with the terminal homology boxes was found to have the capacity to span large distances in terms of chromosomal nucleotide sequences.
3. I-SceI–mediated cleavage of the chromosome followed by intramolecular DSB repair proceeds efficiently. The homologies (box B) involved in the repair are atypically short substrates for RecA; however, activity of RecA is required in the process. Due to the high efficiency of cleavage and repair, replica-plating on antibiotic-containing and nonselective plates to distinguish between unresolved constructs and scarless end products is usually not necessary. Note that I-SceI cleavage not only stimulates intramolecular recombination but also selects for resolved genomic constructs.
4. Cells can be easily cured of the helper plasmids with heat-sensitive replication (pSTKST, pKD46) by growth at nonpermissive temperatures (37°C to 42°C) in the absence of the selective antibiotic.
5. By proper strategy, the time required for introducing multiple changes in the genome can be shortened. Insertion intermediates, marked by a drug resistance cassette, can be made for each modification in parallel cell lines. By repeated sequential application of P1 transduction and DSB-stimulated recombination, these individual modifications can be accumulated in a single strain. In this case, repeated curing and transformation of the helper plasmid pSTKST is not necessary. Instead, by carrying out the steps of P1 transduction and DSB-stimulated recombination at 30°C and by applying Km selection, the plasmid can be maintained continuously in the main cell line.

6. For efficient integration into the chromosome, homology arms of the targeting fragment cloned in the suicide plasmid should be >500 bp long. Longer homologous segments provide better recombination efficiency. However, it must be taken into account that generating homology arms by PCR might inadvertently introduce unwanted point mutations into the genome. It is recommended to use a high-fidelity DNA polymerase and verify the plasmid construct by DNA sequencing.
7. When resolving the suicide plasmid–chromosome cointegrate, the expected ratio of wt and mutant end products is 50 : 50, provided a small mutation is placed between homology arms of equal size. However, several factors can cause deviations from this ratio. When creating a large deletion, the chromosomal distance between the pairs of homology arms is very different. As a consequence, recombination leading to the mutant product usually proceeds less readily, and a large number of colonies must be screened to obtain the desired deletion. Obviously, "one-sided" resolution of the cointegrate leading to reversal to wt is observed when the mutant version of the cell is not viable. A decrease of growth rate, caused by the introduced mutation, can also result in high wt-to-mutant ratio or even in failure to obtain the desired mutant. In such cases, the growth advantage of wt cells can be limited by applying a shorter I-SceI induction period in liquid medium.

Acknowledgments

The authors thank Agnes Szalkanovics for technical assistance in the genome engineering projects and Antal Kiss for critical reading of the manuscript.

References

1. Welch, R. A., Burland, V., Plunkett, G. 3rd, Redford, P., Roesch, P., Rasko, D., et al. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 17020–17024.
2. Pósfai, G., Plunkett, G. 3rd, Fehér, T., Frisch, D., Keil, G. M., Umenhoffer, K., et al. (2006) Emergent properties of reduced-genome *Escherichia coli*. *Science* **312**, 1044–1046.
3. Kolisnychenko, V., Plunkett, G. 3rd, Herring, C. D., Fehér, T., Pósfai, J., Blattner, F. R., and Pósfai, G. (2002) Engineering a reduced *Escherichia coli* genome. *Genome Res.* **12**, 640–647.
4. Pósfai, G., Kolisnychenko, V., Bereczki, Z., and Blattner, F. R. (1999) Markerless gene replacement in *Escherichia coli* stimulated by a double-strand break in the chromosome. *Nucleic Acids Res.* **27**, 4409–4415.
5. Datsenko, K. A., and Wanner, B. L. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6640–6645.
6. Zhang, Y., Buchholz, F., Muyrers, J. P., and Stewart, A. F. (1998) A new logic for DNA engineering using recombination in *Escherichia coli*. *Nat. Genet.* **20**, 123–128.
7. Yu, D., Ellis, H. M., Lee, E. C., Jenkins, N. A., Copeland, N. G., and Court, D. L. (2000) An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5978–5983.
8. Monteilhet, C., Perrin, A., Thierry, A., Colleaux, L., and Dujon, B. (1990) Purification and characterization of the in vitro activity of I-Sce I, a novel and highly specific endonuclease encoded by a group I intron. *Nucleic Acids Res.* **18**, 1407–1413.
9. Serres, M. H., Goswami, S., and Riley, M. (2004) GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins. *Nucleic Acids Res.* **32** (Database issue), D300–302.

10. Suzuki, H., Nishimura, Y., and Hirota, Y. (1978) On the process of cellular division in *Escherichia coli:* a series of mutants of *E. coli* altered in the penicillin-binding proteins. *Proc. Natl. Acad. Sci. U.S.A.* **75**, 664–668.
11. Bernhardt, T. G., and de Boer, P. A. (2004) Screening for synthetic lethal mutants in *Escherichia coli* and identification of EnvC (YibP) as a periplasmic septal ring factor with murein hydrolase activity. *Mol. Microbiol.* **52**, 1255–1269.
12. Pósfai, G., Koob, M. D., Kirkpatrick, H. A., and Blattner, F. R. (1997) Versatile insertion plasmids for targeted genome manipulations in bacteria: isolation, deletion, and rescue of the pathogenicity island LEE of the *Escherichia coli* O157:H7 genome. *J. Bacteriol.* **179**, 4426–4428.
13. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular Cloning. A Laboratory Manual*, 2nd ed. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

17

Minimization of the *Escherichia coli* Genome Using the Tn*5*-Targeted Cre/*loxP* Excision System

Byung Jo Yu and Sun Chang Kim

Summary

Efficient genome-engineering tools have been developed for use in whole-genome essentiality studies. In this chapter, we describe a powerful genomic deletion tool, the Tn*5*-targeted Cre/*loxP* excision system, for determining genetic essentiality and minimizing bacterial genomes on a genome-wide scale. This tool is based on the Tn*5* transposition system, phage P1 transduction, and the Cre/*loxP* excision system. We have generated two large pools of independent transposon insertion mutants in *Escherichia coli* using random transposition of two modified Tn*5* transposons (TnKloxP and TnCloxP) with two different selection markers, kanamycin-resistance gene (Km^R) or chloramphenicol-resistance gene (Cm^R), and a *loxP* site. Transposon integration sites are identified by direct genome sequencing of the genomic DNA. By combining a mapped transposon mutation from each of the mutant pools into the same chromosome using phage P1 transduction and then excising the nonessential genomic regions flanked by the two *loxP* sites using Cre-mediated *loxP* recombination, we can obtain numerous *E. coli* deletion strains from which nonessential regions of the genome are deleted. In addition to the combinatorial deletion of the *E. coli* genomic regions, we can create a cumulative *E. coli* deletion strain from which all the individual deleted regions are excised. This process will eventually yield an *E. coli* strain in which the genome is reduced in size and contains only regions that are essential for viability.

Key Words: combinatorial deletion; Cre/*loxP* excision system; cumulative deletion; genetic essentiality; phage P1 transduction; reduced genome; Tn5 random transposition.

1. Introduction

Bacterial genomes currently are being sequenced at an increasingly rapid rate. Indeed, the complete genome sequences of more than 400 microorganisms have already been determined (*see* Genomes OnLine Database, http://www.genomesonline.org/) and represent a valuable resource for scientists who seek a comprehensive understanding of the cellular life. On the basis of this genome information, researchers in the fields of bacterial genomics and metabolic engineering are in the process of constructing novel bacterial strains with a minimal gene set or a markedly reduced genome and custom-designed microorganisms that produce desired products efficiently (*1–8*). In

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

order to construct these strains, simple and efficient genetic deletion methods are required. However, the existing genetic deletion methods, which include the widely used site-specific recombination systems FLP/*FRT*, Cre/*loxP*, and λ-Red *(9–12)*, require the creation of targeting vectors or complex polymerase chain reaction (PCR) products each time a deletion experiment is performed. Therefore, to delete the many nonessential genes scattered throughout the chromosome, numerous deletion experiments must be performed if one is to generate minimal essential gene sets or construct custom-designed microorganisms. As a consequence, it has become necessary to establish genome-scale procedures that allow scientists to carry out many deletions rapidly, efficiently, and simultaneously.

The Cre/*loxP* excision system is an efficient genomic-deletion method in which Cre, a site-specific recombinase, mediates recombination between DNA sequences that contain a 34-base pair (bp) *loxP* site with high efficiency *(9)*. However, as mentioned above, this method requires the creation of targeting vectors for the insertion of two *loxP* sites into the target regions of a genome whenever a deletion experiment is to be performed. One way to overcome this obstacle is through random transposition of the Tn*5* transposon containing a *loxP* site *(13)*. Tn*5* transposition thus allows the rapid insertion of *loxP* sites into numerous regions of the genome simultaneously. In addition, two *loxP* sites can be brought, in parallel, into a single bacterial strain by bacteriophage P1 transduction *(14)*.

Here, we describe a powerful deletion tool, the Tn*5*-targeted Cre/*loxP* excision system, in which random Tn*5* transposition, phage P1 transduction, and the Cre/*loxP* excision system are combined to achieve the rapid deletion of scattered nonessential genes in *Escherichia coli*. First, for the random insertion of *loxP* sites into the *E. coli* genome, we have constructed two modified Tn*5* transposons, TnKloxP and TnCloxP, which carry the kanamycin-resistance gene (Km^R) and a *loxP* site, and the chloramphenicol-resistance gene (Cm^R) and a *loxP* site, respectively. Both transposons are flanked by the hyperactive 19-bp outer-end transposase recognition sequences (OE) *(13)*. TnKloxP or TnCloxP is mixed with the Tn*5* transposase *in vitro* to form transposase-transposon complexes called transposomes *(5, 13)*. Upon electroporation of these transposomes into the *E. coli* genome, two large pools of independent transposon insertion mutants are generated. The transposon insertion sites are precisely identified by direct genomic sequencing of the genomic DNA.

Depending on the genes to be deleted, we choose a pair of mutant bacterial strains, one from the TnKloxP group and one from the TnCloxP group. Using phage P1 transduction, we constructed Cm^R-Km^R double-resistant strains that contain two *loxP* sites in tandem. Subsequent expression of the Cre recombinase from the pELCre expression plasmid *(9)* induces recombination between the *loxP* sites, resulting in excision of the *loxP*-flanked region of DNA. A scheme summarizing the individual steps of the Tn*5*-targeted Cre/*loxP* excision system is illustrated in **Figure 1**.

Each of the individual deletions are then combined into a single "cumulative deletion strain" using phage P1 transduction. The inserted selection markers (Km^R or Cm^R gene) and a *loxP* site are exchanged with the *FRT*-flanked Tc^R-*sacB* cassette by Red recombination, and the inserted cassette is eliminated by FLP expression *(12)* for further introduction of other deletion regions into the cumulative deletion strain. Continuous

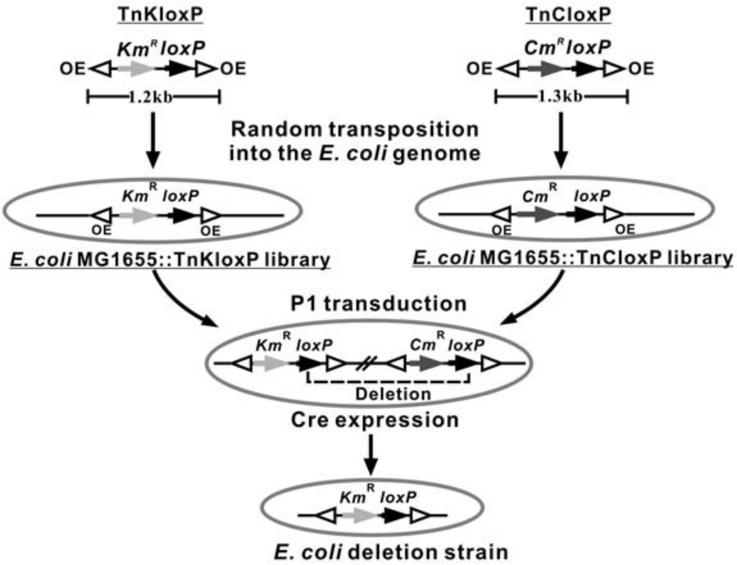


Fig. 1. Overall scheme of the Tn*5*-targeted Cre/*loxP* excision system. Two modified Tn*5*-transposons, TnKloxP and TnCloxP, are introduced into the *E. coli* chromosome randomly, producing two groups of mutant libraries (*E. coli* MG1655::TnKloxP and MG1655::TnCloxP). Two mutant strains with a *loxP* site in the same orientation are selected, one from each mutant library, depending on the target region to be deleted. The selected two *loxP* sites are brought in parallel into a single strain by phage P1 transduction. The selected region between the two *loxP* sites is deleted by the action of the Cre recombinase. OE, outer end transposase recognition sequence; Km^R, kanamycin resistance gene; Cm^R, chloramphenicol resistance gene.

insertions of each deleted region into a single genome result in the generation of an *E. coli* strain in which the genome is reduced in size.

2. Materials

2.1. Bacterial Strains

E. coli K-12 strain MG1655 ***(15)*** is used for genomic deletions, and *E. coli* K-12 DH5α is used for plasmid preparation.

2.2. Plasmids

1. Plasmid pMOD^{TM}⟨MCS⟩, which contains a multiple cloning site (MCS) flanked by the hyperactive 19-bp outer end (OE) sequences that are specifically and uniquely recognized by the EZ-Tn*5*^{TM} transposase, was purchased from Epicentre Biotechnologies (Madison, WI) ***(16)*** (**Note 1**).
2. Plasmids pKKlox and pKClox, which contain Km^R and a *loxP* site, and Cm^R and a *loxP* site, respectively, were obtained from Yoon ***(9)***.
3. Plasmids pTnKloxP and pTnCloxP are constructed for the preparation of the TnKeoxP and TnCloxP transposons (**Fig. 2A**).
4. Plasmid pELCre was constructed by cloning a DNA fragment that contained a *tetR-*P_{tet}*-cre*, which was generated by PCR from pTATCπ ***(9)***, into the *Bam*HI site of pEL3 ***(17)***. The

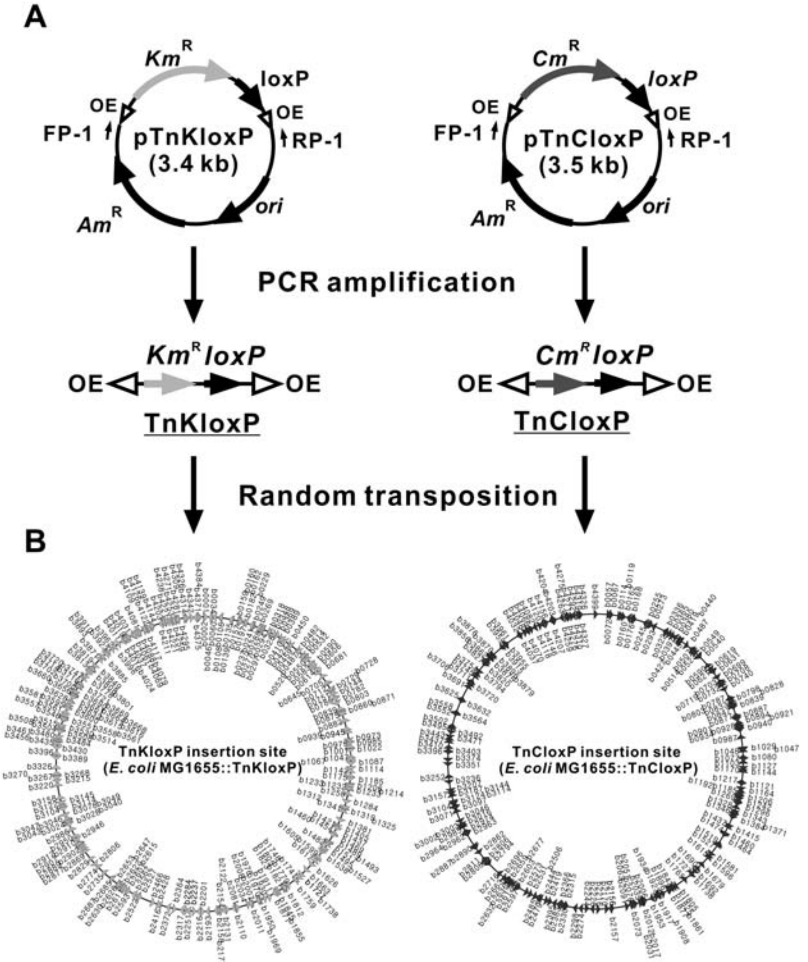


Fig. 2. Generation of TnKloxP and TnCloxP **(A)** and physical maps **(B)** of the inserted TnKloxP and TnCloxP in the *E. coli* genome. **(A)** The linear transposons TnKloxP and TnCloxP are generated from pTnKloxP and pTnCloxP, respectively, by PCR amplification using pMOD⟨MCS⟩ primers FP-1(5′-ATTCAGGCTGCGCAACTGT-3′) and RP-1 (5′-TCAGT GAGCGAGGAAGCGGAAG-3′). TnKloxP consists of *Km*^{R} and a *loxP* site, and TnCloxP consists of *Cm*^{R} and a *loxP* site. Both transposons are flanked by the hyperactive 19-bp outer end transposase recognition sequence (OE:5′-CTGTCTCTTATACACATCT-3′). **(B)** Physical maps of *E. coli* MG1655 chromosome targeted with TnKloxP and with TnCloxP. The arrows in the *E. coli* MG1655::TnKloxP (gray) and MG1655::TnCloxP (black) libraries indicate the insertion sites of TnKloxP and TnCloxP. The positions of the inserted transposons are indicated using Blattner numbers. The Blattner numbers located inside and outside the circle represent leftward and rightward orientations of the inserted *loxP* site, respectively. Information concerning the bacterial strains used in these libraries is also available on our Web site (http://bio.kaist.ac.kr/~mbtlab/).

tetR-P$_{tet}$-cre DNA fragment contains a tetracycline-repressor gene and the *cre* gene driven by a tetracycline-responsive promoter.

5. Plasmids pKD4, pKD46, and pCP20 were obtained from Barry Wanner ***(12)*** and used to eliminate selection markers that remained after the replacement of deletion targets.
6. pST is a derivative of pKD4 that contains the *FRT*-flanked *Bacillus subtilis* levansucrase gene (*sacB*) and the tetracycline resistance gene (Tc^R).

2.3. Media, Solution, and Enzymes

1. Luria-Bertani (LB) medium: 1% bacto-tryptone, 0.5% yeast extract, and 0.5% NaCl, sterilized by autoclaving.
2. LB plates: LB medium supplemented before autoclaving with 1.5% agar. R top agar plates: 1% bacto-tryptone, 0.1% bacto yeast extract, 0.8% Difco bacto agar, 0.8% NaCl, 2 mM $CaCl_2$, and 0.1% glucose. R plate: 1% bacto-tryptone, 0.1% bacto yeast extract, 1.2% Difco bacto agar, 0.8% NaCl, 2 mM $CaCl_2$, and 0.1% glucose.
3. Ampicillin (Ap: 50 µg/mL).
4. Kanamycin (Km: 25 µg/mL).
5. Chloramphenicol (Cm: 17 µg/mL).
6. Tetracycline (Tc: 25 µg/mL).
7. Heat-inactivated chlortetracycline (cTc: 15 µg/mL).
8. Glycerol.
9. MC buffer: 10 mM $MgSO_4$ and 5 mM $CaCl_2$.
10. 1 M sodium citrate.
11. Chloroform.
12. Enzymes: *Bam*HI, *Not*I, *Xba*I, alkaline phosphatase, T4 DNA ligase (New England Biolabs, Beverly, MA), Tn*5* transposase (1 U/µL) (Epicentre Biotechnologies), and ExTaq polymerase (Takara, Japan).
13. Wizard Plus SV minipreps DNA purification system (Promega, Madison, WI).
14. Genomic DNA purification system (Qiagen, Hilden, Germany).
15. Master Pure DNA isolation Kit (Epicentre Biotechnologies).
16. Gel electrophoresis solutions and reagents: agarose, ethidium bromide (EtBr: 500 µg/mL), and Tris-borate-EDTA (TBE) electrophoresis buffer (45 mM Tri-borate and 1 mM EDTA).

2.4. Laboratory Equipment

1. Gene Pulser system (Bio-Rad, Hercules, CA).
2. Spectrophotometer.
3. Gel electrophoresis apparatus and equipment.
4. Power supplies.
5. Water bath.
6. Environmental incubators.
7. Autoclave.
8. Southern blot hybridization equipment.

2.5. Oligonucleotides and Direct Sequencing

Primers FP-1(5′-ATTCAGGCTGCGCAACTGT-3′), RP–1(5′-TCAGTGAGCGAG GAAGCG GAAG-3′), Tn5Int (5′-TCGACCTGCAGGCATGCAAGCTTCA-3′), and locus-specific primers were synthesized by GenoTech (Daejeon, Korea). Direct sequencing of mutant genomes was conducted by SolGent (Daejeon, Korea).

3. Methods

The methods described below outline the preparation of two transposons, TnKloxP and TnCloxP; construction of transposon insertion libraries; combinatorial deletion of *E. coli* genomic regions; and cumulative deletion of *E. coli* genomic regions to generate a reduced-size genome.

3.1. Construction of TnKloxP and TnCloxP and Production of the Corresponding Transposomes

In this section, we describe the construction of (a) plasmids pTnKloxP and pTnCloxP, (b) transposons TnKloxP and TnCloxP, and (c) the TnKloxP and TnCloxP transposomes.

3.1.1. Construction of the pTnKloxP and pTnCloxP Plasmids

For the construction of pTnKloxP, a 1.1-kb DNA fragment that contains *Km*^{R} and a *loxP* site is isolated by digesting pKKlox with *Not*I and *Xba*I and is cloned into the *Bam*HI site of pMOD^{TM}⟨MCS⟩ by blunt-end ligation. pTnCloxP is also constructed by inserting a 1.2-kb DNA fragment that contains *Cm*^{R} and a *loxP* site, which can be obtained by digesting pKClox with *Not*I and *Bam*HI, into the *Bam*HI site of pMOD^{TM}⟨MCS⟩ by blunt-end ligation.

3.1.2. Preparation of the TnKloxP and TnCloxP Transposon DNA

The linear transposons TnKloxP and TnCloxP are generated from plasmids pTnKloxP and pTnCloxP, respectively, by PCR amplification using primers pMOD⟨MCS⟩ FP-1(5′-ATTCAGGCTGCGCAACTGT-3′) and pMOD⟨MCS⟩ RP–1 (5′-TCAGT GAGCGAGGAAGCGGAAG-3′) (**Fig. 2A**). The PCR is carried out for 25 cycles (30 s at 94°C; 30 s at 50°C; 90 s at 72°C). TnKloxP contains *Km*^{R} and a *loxP* site, and TnCloxP contains *Cm*^{R} and a *loxP* site. Both transposons are flanked by the hyperactive 19-bp outer-end transposase recognition sequence (OE:5′-CTGTCTCTTATACA CATCT-3′). The PCR-amplified TnKloxP and TnCloxP are purified with the Qiagen PCR purification kit and suspended in 50 µL of nuclease-free water.

3.1.3. Production of the TnKloxP and TnCloxP Transposomes

To produce transposase-transposon complexes called transposomes, purified TnKloxP and TnCloxP DNA is incubated with Tn*5* transposase *in vitro*. Production of stable transposomes can only be accomplished in the absence of Mg^{2+}. The detailed procedure is as follows:

1. Prepare the transposome reaction mixture by adding in the following order:
   2 µL TnKloxP or TnCloxP (250 µg/mL in distilled, deionized water [DDW])
   4 µL transposase (1 U/µL)
   2 µL 100% glycerol
   2 µL DDW
2. Mix by vortexing and incubate for 30 min at room temperature.
3. Store the solution in aliquots at –20°C. The solution will not freeze and is stable for at least 1 year.
4. Use 1 µL of the transposome for electroporation into competent *E. coli* MG1655 cells.

3.2. Construction of Transposon Insertion Libraries

Construction of transposon-insertion libraries is described in **Section 3.2.1** to **Section 3.2.3**. This step includes the random transposition of two kinds of transposomes (TnKloxP and TnCloxP) into the *E. coli* MG1655 genome, the confirmation of random transposition of the transposons using Southern blot analysis, and the identification of transposon insertion locations by direct genomic sequencing.

3.2.1. Random Transposition of TnKloxP and TnCloxP Transposomes into the E. coli *Genome*

Electroporation of the transposomes are performed using a standard procedure ***(16)***, as follows:

1. Grow *E. coli* MG1655 in 100 mL LB to mid-log phase ($OD_{600} = 0.5$) at 37°C.
2. To render the bacterial cells electrocompetent, chill the cells, harvest them by centrifugation, wash the cell pellets with ice-cold water two times and 10% glycerol one time, and suspend the cells in 150 µL 10% glycerol.
3. Add 1 µL of the transposome solution (TnKloxP or TnCloxP) to 50 µL of the electrocompetent *E. coli* cells and transfer the mixture to a 2.0-mm gap electroporation cuvette (Bio-Rad).
4. Electroporate the mixture at 2.5 kV, 25 µFD, and 200 Ω.
5. Dilute the electroporated cells to 1 mL with LB and incubate the mixture for 1 h at 37°C with agitation. The electroporated transposomes are activated by Mg^{2+} present inside cells, and then their transposon components are inserted randomly into the *E. coli* MG1655 genome ***(6)***.
6. Spread the cells electrotransformed with TnKloxP or TnCloxP on LB plates containing Km or Cm, respectively, and incubate overnight at 37°C.
7. Select single colonies and grow overnight at 37°C on LB plates containing either Km or Cm.

3.2.2. Confirmation of Random Transposition by Southern Blot Analysis

Random transpositions of TnKloxP and TnCloxP into the *E. coli* genome are confirmed by Southern blot analysis performed under the following conditions:

1. Isolate genomic DNA from the transposon insertion mutants with the Genomic DNA Purification Kit (Qiagen).
2. Digest the isolated genomic DNA (10 µg) with *Cla*I by incubating the DNA with the enzyme for 16 h at 37°C.
3. Separate the resulting DNA fragments by electrophoresis through a 0.8% agarose gel at 40 V to allow the DNA to migrate slowly.
4. Transfer the gel containing the DNA fragments to a Hybond N+ membrane (Amersham Biosciences, Piscataway, NJ) and hybridize the membrane with ^{32}P-labeled DNA fragments that correspond with the selection markers (Km^R or Cm^R gene) of the transposons, as described in the standard procedure ***(18)***.
5. To obtain an autoradiographic image, cover the membrane with a sheet of clear plastic wrap and expose the membrane to X-ray film for 16 to 24 h at –70°C with an intensifying screen.

Under our experimental conditions, Southern blot analysis should show that all the selected mutants have only one transposon insertion in each *E. coli* genome (**Fig. 3**).

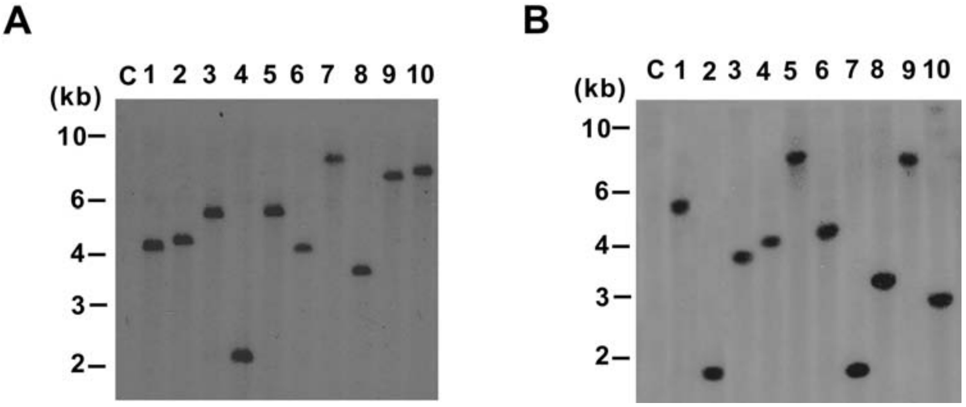


Fig. 3. Confirmation of random transposition events using Southern blot hybridization analysis. Isolated genomic DNA from *E. coli* MG1655 (lane C) and the transposon insertion strains (lanes 1 to 10) is digested with *Cla*I, and ^{32}P-labeled DNA fragments that correspond with the selection markers, *Km*^{R} **(A)** or *Cm*^{R} **(B)** of the transposons are used as probes, respectively. Southern blot analysis shows that all the selected mutants have only one transposon insertion in the chromosome under our experimental conditions.

3.2.3. Identification of the Transposon Insertion Locations by Direct Genomic Sequencing

The genomic DNA of all selected mutants is subjected to direct sequencing as follows:

1. Isolate genomic DNA from the individual transposon insertion mutants using the Master Pure DNA Isolation Kit (Epicentre Biotechnologies) (**Note 2**).
2. Add the isolated genomic DNA (2 µg) and the Tn*5*Int primer (5′-TCGACCTGCAGGCAT GCAAGCTTCA-3′) (25 pmol) to 50-µL reactions in which is contained 16 µL of dye terminator premix of the BigDye Terminator Cycle Sequencing Kit (PE Biosystems, Foster City, CA).
3. Perform thermocycling of the sequencing reaction according to the following program: one cycle of 4 min at 95°C, and 60 cycles of 30 s at 95°C and 4 min at 60°C.
4. Purify the amplified products and electrophorese the products in an ABI 310 Genetic Analyzer (Applied Biosystems, Foster City, CA) with ABI version 3.3 Sequence Analysis software.
5. Compare the sequences with the GeneBank DNA sequence database using the BLAST program (http://www.ncbi.nlm.nih.gov/BLAST/). The location and direction of the inserted *loxP* sites in the mutant strains are shown on the physical map of the *E. coli* genome in **Figure 2B**, and the information concerning the bacterial strains in these mutant libraries is also available on our Web site (http://bio.kaist.ac.kr/~mbtlab/).

3.3 Combinatorial Deletion of Various Regions of E. coli Genomic DNA

Combinatorial deletions of various regions of *E. coli* genomic DNA are described in **Section 3.3.1** to **Section 3.3.3**. This section includes information on (a) the construction of *Cm*^{R}-*Km*^{R} double-resistant strains with two *loxP* sites in tandem using phage P1 transduction, (b) the expression of Cre recombinase for deleting targeted genomic regions flanked by two *loxP* sites, and (c) the selection of genomic deletion mutants (**Fig. 4**).

3.3.1. Insertion of TnKloxP and TnCloxP into the Same E. coli *Genome Using Phage P1 Transduction*

Depending on the genes to be deleted, we choose a pair of mutant *E. coli* strains, one from the TnKloxP insertion mutant library and one from the TnCloxP insertion mutant library. Using phage P1 transduction, we construct Cm^R-Km^R double-resistant strains that contain two *loxP* sites in tandem. The position and direction of the two *loxP* sites and the possible presence of essential genes in the region flanked by the two *loxP* sites should be considered (**Note 3**).

To construct mutant strains that contain two *loxP* sites flanking a target region of DNA, phage P1 transduction is performed as follows:

3.3.1.1. Part 1: Preparation of the Phage P1 Lysate

Prepare phage P1 lysates with mutant *E. coli* strains that carry TnKloxP (donor cells) and use them for the transduction of the mutant *E. coli* strains that carry TnCloxP (recipient cells) or vice versa.

1. Grow the donor cells (TnKloxP or TnCloxP insertion mutants) to stationary phase in 3 mL LB broth.
2. Transfer the culture to fresh LB (3 mL) that contains 5 mM $CaCl_2$ and incubate at 37°C to an OD_{600} = 0.4.

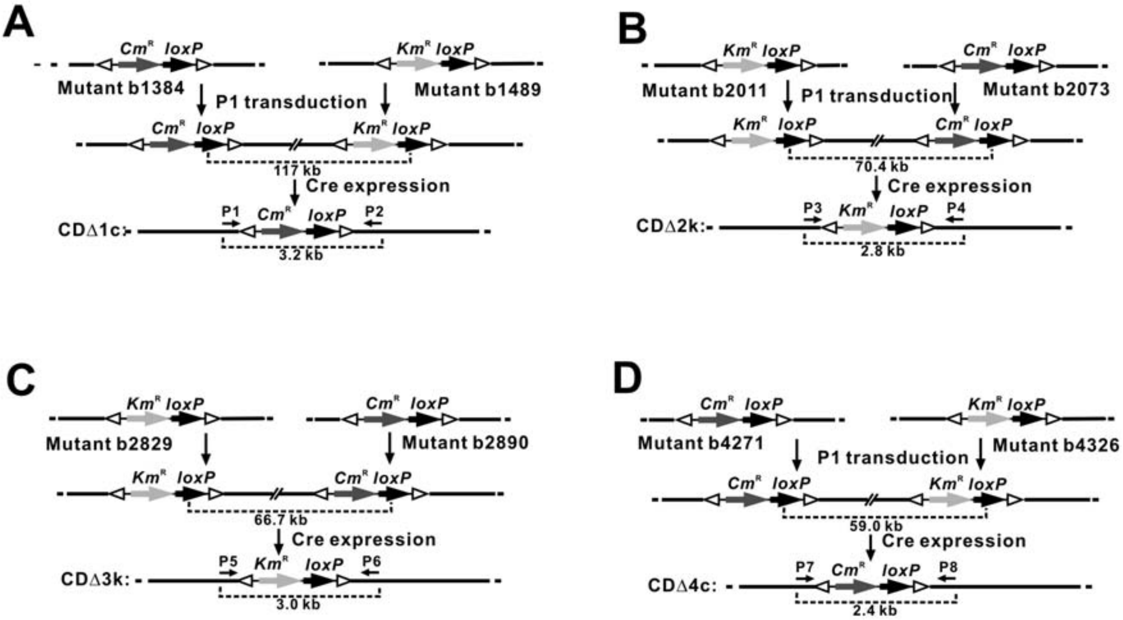


Fig. 4. Combinatorial deletions. For the generation of combinatorial deletions, two *loxP* sites are chosen, one from each Tn*5*-mutant library, and recombined in tandem on the same chromosome by phage P1 transduction. After Cre expression, the selected regions flanked by the two *loxP* sites are deleted, producing deletion strains CDΔ1c (**A**), CDΔ2k (**B**), CDΔ3k (**C**), and CDΔ4c (**D**). The deletions of CDΔ1c, CDΔ2k, CDΔ3k, and CDΔ4c are confirmed by PCRs using locus-specific primers (P1 to P8).

3. Add 2 µL of phage P1 lysate (10^7 phages) to 1 mL of donor cells and incubate the mixture for 20 min at 37°C without agitation.
4. Mix the P1-infected cells with 5 mL R-top agar, pour the mixture into the R plate, and incubate overnight at 37°C.
5. Scrape the soft agar layer and transfer it to a 15-mL tube.
6. Add 500 µL $CHCl_3$, vortex, and centrifuge for 15 min at 3000 rpm.
7. Collect the supernatant and use as phage P1 lysates.

3.3.1.2. Part 2: Transduction of *E. coli* Recipient Cells with the Phage P1 Prepared in Part 1

1. Grow the recipient strain (TnCloxP or TnKloxP insertion mutants) to stationary phase in 3 mL LB.
2. Centrifuge the cells at room temperature for 5 min at 1500 × g.
3. Suspend the cell pellet in 1 mL MC buffer (10 mM $MgSO_4$ and 5 mM $CaCl_2$).
4. Add 100 µL of the phage P1 preparation diluted 10- or 100-fold in phage dilution buffer to a series of Eppendorf tubes.
5. Add 100 µL of the concentrated recipient cells to the Eppendorf tubes. As controls, prepare one tube with recipient cells and no phage, and another with concentrated phage and no recipient cells. Incubate the tubes for 20 min at 37°C without agitation.
6. After the incubation, add 100 µL of 1 M sodium citrate to each tube. This chelates the Ca^{2+} and Mg^{2+} required for phage adsorption.
7. Centrifuge the cells for 30 s at room temperature.
8. Suspend the cells in 100 µL of 100 mM sodium citrate and then centrifuge the cells for 30 s at room temperature.
9. Suspend the cells in 1 mL LB and incubate the cells for 45 min at 37°C.
10. Spread the cells on LB plates containing both Km and Cm and incubate overnight at 37°C.

The *Cm*^{R}-*Km*^{R} double-resistant strains that contain two *loxP* sites in parallel are selected on the LB plates supplemented with Cm and Km.

3.3.2. Cre Expression

For the expression of Cre in the *Cm*^{R}-*Km*^{R} double-resistant strains of *E. coli*, the strains are transformed with the Cre expression plasmid (pELCre). Induction of Cre expression by incubation with cTc induces a *loxP*-mediated recombination event, resulting in deletion of the genomic region flanked by the two *loxP* sites and leaving behind a single *loxP* site and a selection marker (*Cm*^{R} or *Km*^{R}) at the site of the deleted target genomic region. Transformation and induction are accomplished as follows:

1. Prepare the electrocompetent *Cm*^{R}-*Km*^{R} double-resistant cells as described in **Section 3.2.1**.
2. Transform pELCre into the competent cells by electroporation and incubate the electroporated cells in 1 mL LB on a rotary shaker for 1 h at 30°C.
3. Spread the pELCre-transformed cells on LB plates containing Ap and incubate overnight at 30°C.
4. Select single colonies, inoculate into 3 mL LB containing Ap and cTc, and incubate overnight on a rotary shaker at 30°C.

5. Streak out a small aliquot (2 μL) of the cultures on LB replica plates, first on LB plates containing Cm and then on LB plates containing Km, or vice versa.
6. Cure pELCre by shifting the culture temperature to 42°C.

3.3.3. Verification of Correct Genomic Deletions

With the Tn*5*-targeted Cre/*loxP* excision system described above, four individual genomic regions described in **Figure 4**—deletion region (DR)1c:b1384-b1489, DR2k: b2011-b2073, DR3k:b2829-b2890, and DR4c:b4271-b4326—are deleted, producing chromosomal deletion (CD) strains CDΔ1c, CDΔ2k, CDΔ3k, and CDΔ4c. PCR with locus-specific primers (P1 to P8) is carried out to confirm that the deletion from the *E. coli* genome is performed correctly. The PCRs are performed using a pair of primers that are specific to the end points of each deletion (P1-P2 for CDΔ1c, P3-P4 for CDΔ2k, P5-P6 for CDΔ3k, and P7-P8 for CDΔ4c) (**Fig. 4**). The successful deletions of CDΔ1c, CDΔ2k, CDΔ3k, and CDΔ4c produce PCR products of 3.2, 2.8, 3.0, and 2.4 kb, respectively (*see* **Fig. 6A**). In addition, the absence of each deleted region from the deletion strains is confirmed by multiplex PCRs with two pairs of primers specific to internal genes of each deletion region (the selected internal genes are: b1407 and b1468 for DR1, b2035 and b2060 for DR2, b2836 and b2870 for DR3, and b4291 and b4317 for DR4, respectively) (*see* **Fig. 6B**) (**Note 4**). We carry out these PCRs using genomic DNAs from *E. coli* MG1655 (wild type) and the deletion strains (CDΔ1c, CDΔ2k, CDΔ3k, and CDΔ4c) as the templates.

3.4. Cumulative Deletion of* E. coli *Genomic Regions

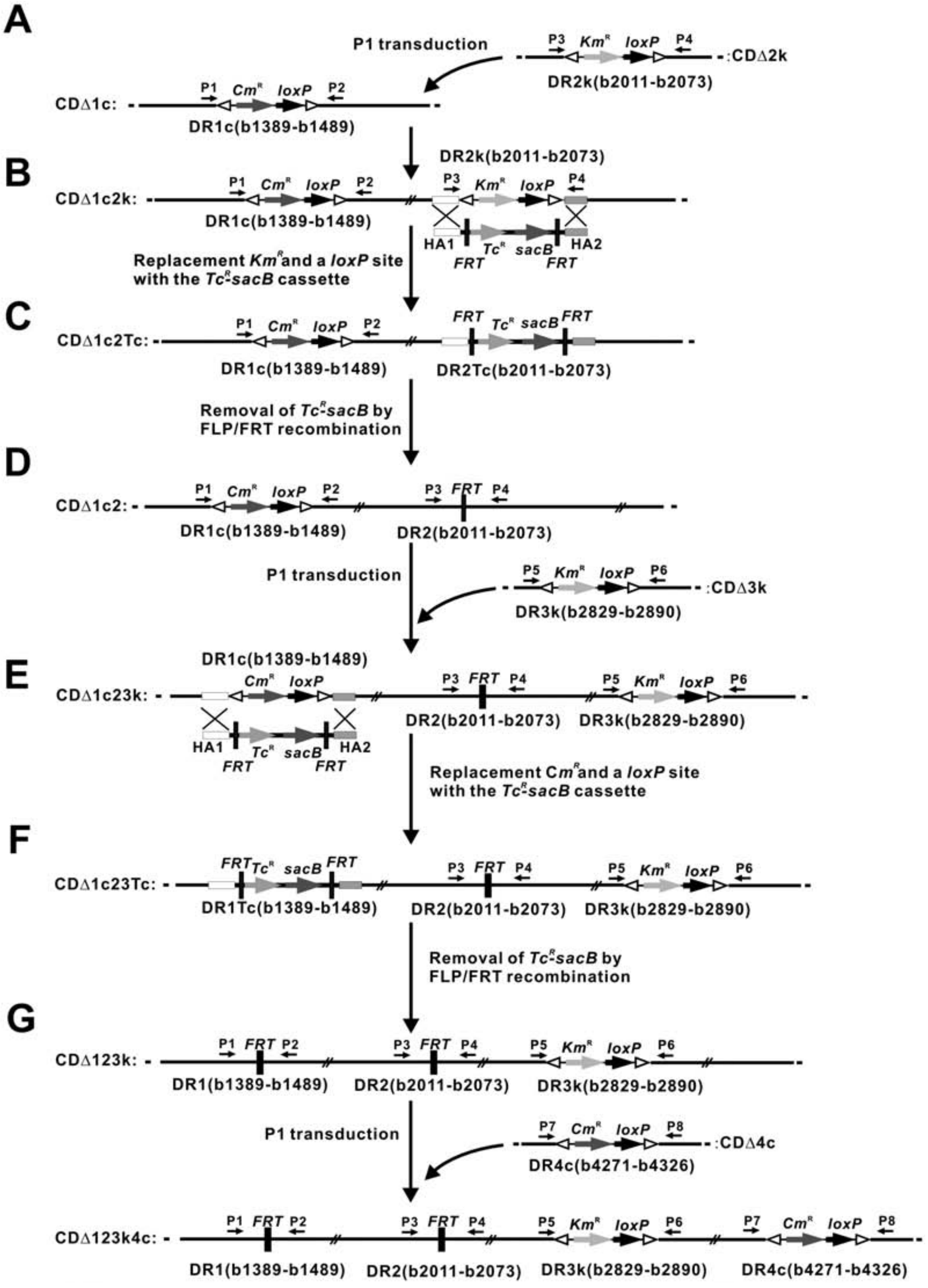
Next, we generate a cumulative deletion *E. coli* strain (CDΔ123k4c) as described in **Section 3.4.1** to **Section 3.4.4**. This section includes information on (a) the cumulative deletion in which the DR2k of CDΔ2k is brought into the CDΔ1c using phage P1 transduction, generating CDΔ1c2k, (b) replacement of the selection marker *Km^R* and a *loxP* site present on the DR2k of CDΔ1c2k with the *FRT*-flanked *Tc^R-sacB* cassette and removal of the inserted selection marker *Tc^R-sacB* by FLP expression, producing *Km^R*-free CDΔ1c2 strain, and (c) transfer of DR3k into CDΔ1c2 to produce CDΔ1c23k, and subsequent transfer of DR4c into CDΔ123k, resulting in the construction of a cumulative strain CDΔ123k4c (**Fig. 5**). The cumulative deletion strain, CDΔ123k4c is prepared as follows.

3.4.1. Construction of CDΔ1c2k Using Phage P1 Transduction

1. Select the CDΔ1c to serve as phage P1 recipient cells and the CDΔ2k to serve as donor cells (**Fig. 5A**).
2. Transfer DR2k of CDΔ2k into the CDΔ1c by using phage P1 transduction (**Section 3.3.1**).
3. Select the CDΔ1c2k on LB plates that contain both Km and Cm and verify the deletion regions by using the PCR analyses as described in **Section 3.3.3**.

3.4.2. Replacement of the Selection Marker Km^R *Gene and a* loxP *Site of CDΔ1c2k with the* FRT*-Flanked* Tc^R-sacB *Cassettes*

Before further introduction of DR3k(b2829-b2890) of CDΔ3k into the genome of CDΔ1c2k, the *Km^R* and a *loxP* site present in DR2k of CDΔ1c2k is exchanged with the *FRT*-flanked *Tc^R-sacB* cassette as follows:


A
P1 transduction
P3 KmR loxP P4 :CDΔ2k
DR2k(b2011-b2073)
CDΔ1c: P1 CmR loxP P2
DR1c(b1389-b1489)
B
DR2k(b2011-b2073)
CDΔ1c2k: P1 CmR loxP P2 P3 KmR loxP P4
DR1c(b1389-b1489)
HA1 TcR sacB HA2
FRT FRT
Replacement KmR and a loxP site
with the TcR-sacB cassette
C
CDΔ1c2Tc: P1 CmR loxP P2 FRT TcR sacB FRT
DR1c(b1389-b1489)
DR2Tc(b2011-b2073)
Removal of TcR-sacB by
FLP/FRT recombination
D
CDΔ1c2: P1 CmR loxP P2 P3 FRT P4
DR1c(b1389-b1489)
DR2(b2011-b2073)
P1 transduction
P5 KmR loxP P6 :CDΔ3k
DR3k(b2829-b2890)
E
DR1c(b1389-b1489)
CDΔ1c23k: CmR loxP P3 FRT P4 P5 KmR loxP P6
HA1 TcR sacB HA2
FRT FRT
DR2(b2011-b2073) DR3k(b2829-b2890)
Replacement CmR and a loxP site
with the TcR-sacB cassette
F
CDΔ1c23Tc: FRT TcR sacB FRT P3 FRT P4 P5 KmR loxP P6
DR1Tc(b1389-b1489) DR2(b2011-b2073) DR3k(b2829-b2890)
Removal of TcR-sacB by
FLP/FRT recombination
G
CDΔ123k: P1 FRT P2 P3 FRT P4 P5 KmR loxP P6
DR1(b1389-b1489) DR2(b2011-b2073) DR3k(b2829-b2890)
P1 transduction
P7 CmR loxP P8 :CDΔ4c
DR4c(b4271-b4326)
CDΔ123k4c: P1 FRT P2 P3 FRT P4 P5 KmR loxP P6 P7 CmR loxP P8
DR1(b1389-b1489) DR2(b2011-b2073) DR3k(b2829-b2890) DR4c(b4271-b4326)

1. Prepare electrocompetent CDΔ1c2k cells, transform pKD46 into the electrocompetent cells, and select transformants harboring pKD46 on LB plates containing Ap.
2. Prepare electrocompetent CDΔ1c2k cells harboring pKD46.
3. Prepare the *FRT*-flanked Tc^R-*sacB* cassette that can replace the Km^R and a *loxP* site present on the DR2k of CDΔ1c2k (each cassette can be generated from pST by PCR using primers with 50-bp extensions homologous to DNA regions flanking the deleted region *[12]*) (**Note 5**).
4. Electrotransform the *FRT*-flanked Tc^R-*sacB* cassette into the electrocompetent CDΔ1c2k cells harboring pKD46, to replace a Km^R and a *loxP* site of DR2k with the cassette (**Fig. 5B**).
5. Select the CDΔ1c2Tc strains on LB plates containing both Cm and Tc.
6. Cure pKD46 from the CDΔ1c2Tc cells harboring pKD46 by shifting to 42°C.
7. Prepare electrocompetent CDΔ1c2Tc cells and transform pCP20 (the FLP expression vector) into the electrocompetent cells.
8. Select transformants on LB plate containing Ap at 30°C.
9. Incubate the CDΔ1c2Tc cells harboring pCP20 in 3 mL of LB at 42°C and streak out a small aliquot (2 µL) from the culture onto LB plates containing Cm. (In this step, pCP20 expresses FLP that can induce an FLP/*FRT* recombination event, eliminating the inserted selection marker Tc^R-*sacB*, and the pCP20 is cured by shifting the incubation temperature to 42°C [**Fig. 5C**]).
10. Select the CDΔ1c2, Km^R-Tc^R-Ap^R free strains, by replica plating on LB plates containing Cm and then LB plates containing Km, Ap, and Tc.

3.4.3. Transfer of DR3k into CDΔ1c2 to Produce CDΔ1c23k, and Subsequent Transfer of DR4c into CDΔ123k for the Construction of CDΔ123k4c

1. Transfer the DR3k of CDΔ3k into the CDΔ1c2 by phage P1 transduction, producing CDΔ1c23k (**Fig. 5D**).
2. Replace the Cm^R and a *loxP* site present in the CDΔ1c23k with the *FRT*-flanked Tc^R-*sacB* cassettes, remove the inserted selection marker Tc^R-*sacB* as described in **Section 3.4.2**, and select Cm^R-free CDΔ123k strains on LB plates containing Km (**Fig. 5E, F**).
3. Transfer CDΔ4c into the CDΔ123k by phage P1 transduction and select CDΔ123k4c strains on LB plates containing Km and Cm (**Fig. 5G**).

Fig. 5. Cumulative deletions. For the generation of a cumulative deletion strain, the deletion mutants CDΔ1c and CDΔ2k are selected as a phage P1 recipient and donor cell, respectively. Phage P1 transduction results in the generation of a deletion strain, CDΔ1c2k, that lacked two genomic regions, deletion region (DR)1c(b1384-b1489) and DR2k(b2011-b2073) **(A)**. Before further introduction of DR3k(b2829-b2890) of CDΔ3k into the genome of CDΔ1c2k, the Km^R and a *loxP* site present in DR2k of CDΔ1c2k is exchanged with the *FRT*-flanked Tc^R-*sacB* cassette **(B)**. The inserted cassette is also eliminated by FLP expression **(C)**. A deletion strain CDΔ1c23k, which contains three deleted genomic regions, DR1c(b1384-b1489), DR2(b2011-b2073), and DR3k(b2829-b2890), is constructed by subjecting strain CDΔ3k to phage P1 transduction **(D)**. Next, removal of the Cm^R and a *loxP* site present in DR1c of CDΔ1c23k **(E, F)** and subsequent introduction of CDΔ4c result in the generation of a cumulative deletion strain CDΔ123k4c, in which four genomic regions, DR1(b1384-b1489), DR2(b2011-b2073), DR3k(b2829-b2890), and DR4c(b4271-b4326), are completely deleted **(G)**.

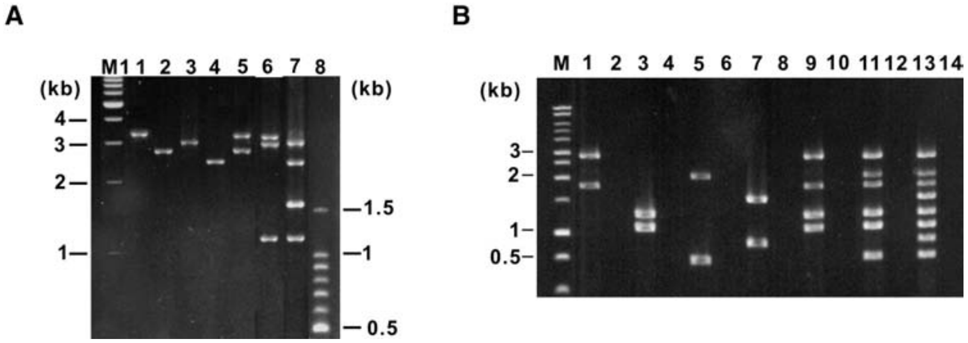


Fig. 6. Verification of genomic deletions in the combinatorial and cumulative deletion strains using PCR analyses. **(A)** The PCR amplifications are performed using a pair of primers specific to the end points of each deletion (P1-P2 for CDΔ1c, P3-P4 for CDΔ2k, P5-P6 for CDΔ3k, and P7-P8 for CDΔ4c, respectively, as indicated in **Fig. 4**). The successful deletion of CDΔ1c, CDΔ2k, CDΔ3k, and CDΔ4c produced PCR products of 3.2, 2.8, 3.0, and 2.4 kb, respectively. Lanes 1 to 4 show the PCR results obtained with the combinatorial deletion strains CDΔ1c, CDΔ2k, CDΔ3k, and CDΔ4c, respectively. Lane 5 shows the multiplex-PCR result obtained with the cumulative deletion strain CDΔ1c2k**.** Lane 6 displays the multiplex-PCR result obtained with the cumulative deletion strain CDΔ1c23k. Lane 7 shows the multiplex-PCR result obtained with the cumulative deletion strain CDΔ123k4c. **(B)** Verification of the absence of deletion region (DR) from each deletion strain. Absence of each DR in deletion strains is confirmed by multiplex PCRs using two pairs of primers specific to the internal sites (genes) of each deletion region (internal genes selected for primers: b1407 and b1468 for DR1, b2035 and b2060 for DR2, b2836 and b2870 for DR3, and b4291 and b4317 for DR4, respectively). Lanes 1, 3, 5, 7, 9, 11, and 13 show the multiplex-PCR results obtained with *E. coli* MGl655. Lanes 2, 4, 6, 8, 10, 12, and 14 indicate the multiplex-PCR results obtained with deletion strains CDΔ1c, CDΔ2k, CDΔ3k, CDΔ4c, CDΔ1c2k, CDΔ1c23k, and CDΔ123k4c, respectively. M, molecular size markers.

4. Confirm the complete and precise removal of genomic regions targeted by each deletion with multiplex PCRs using locus-specific primers as described in **Section 3.3.3** (**Fig. 6**) (**Note 6**).

Using the Tn*5*-targeted Cre/*loxP* excision system, we can construct a single “cumulative deletion strain” (CDΔ123k4c) in which a total of 108 open-reading frames (ORFs) of known function and 179 ORFs of unknown function have been deleted, but which still exhibits a normal growth under standard laboratory conditions *(5)*. This cumulative deletion procedure can be repeated to further minimize the *E. coli* genome to the extent of individual research needs.

4. Conclusion

The major advantage of the Tn*5*-targeted Cre/*loxP* excision system described here is that we can target almost any nonessential region of the *E. coli* genome and delete all the targeted genomic regions simultaneously as long as we have two transposon-saturated *E. coli* libraries in which most of the genomic regions are targeted with a *loxP* site. In addition, we can perform numerous combinatorial and cumulative deletions of the *E. coli* genome by using the *E. coli* transposon insertion mutant libraries.

This practice would allow researchers to circumvent some of the time-consuming steps in the process of generating a minimized genome and custom-designed microorganisms.

Therefore, the Tn*5*-targeted Cre/*loxP* excision system is an extremely powerful tool, not only for the construction of numerous *E. coli* deletion mutants, but also for the functional study of the *E. coli* genome.

Notes

1. Various kinds of transposons can be constructed using pMOD⟨MCS⟩ vector (Epicentre Biotechnologies), which contains a multiple cloning site between 19-bp OE sequences. Various markers that are selectable in each of the target organisms and other desired genetic elements can be cloned into pMOD⟨MCS⟩.
2. To prevent poor direct sequencing results, it is important to obtain high-quality genomic DNA.
3. The directions and locations of the inserted transposons should be considered before conducting phage P1 transduction. This is because half the inserted transposons will insert into the genome in the reverse orientation, and recombination between two oppositely oriented *loxP* sites can induce inversion of the target region throughout the recombination.
4. It is possible that the excised genomic regions can be reintegrated into another location in the genome after Cre/*loxP* recombination. In addition, phage P1 can transfer the complementary genomic region that corresponds with a deleted genomic region during the cumulative deletion. Therefore, complete removal of the deletion regions should be confirmed at every step by multiplex PCRs with primers specific to the internal sites of all deleted regions.
5. During the cumulative deletion, homologous regions-flanked [Km^R (Cm^R)-*sacB*-I-*Sce*I site] cassette, instead of the *FRT*-flanked Tc^R-*sacB* cassette, can be used for the clean removal of the inserted selection marker and a *loxP* site on each deletion region using a double-strand break (DSB)-mediated intramolecular recombination induced by meganuclease I-*Sce*I and the *sacB*/sucrose counter-selection system *(21)*.
6. Because most transposons insert into the internal region of a gene, truncated genes or hybrid genes can be formed after genomic deletion. In addition, genomic rearrangements might occur as a result of the FLP-promoted recombination event between *FRT* sites at different loci. Although such events are rare, it is necessary to confirm all deletion regions by PCR analyses in a cumulative deletion strain.

Acknowledgments

This work was supported by grants from the 21C Frontier Program of Microbial Genomics and Applications (MG05-0204-1-0); the Molecular and Cellular BioDiscovery Research Program (M1-0106-00-0200) from the Ministry of Science and Technology of Korea; the Korea Science and Engineering Foundation (R01-2005-000-11010-0); and the Korea Research Foundation (KRF-2004-042-D00072).

References

1. Hashimoto, M., Ichimura, T., Mizoguchi, H., Tanaka, K., Fujimitsu, K., Keyamura, K., et al. (2005) Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. *Mol. Microbiol.* **55**, 137–149.

2. Kolisnychenko, V., Plunkett, G. 3rd, Herring, C. D., Feher, T., Posfai, J., Blattner, F. R., and Posfai, G. (2002) Engineering a reduced *Escherichia coli* genome. *Genome Res.* **12**, 640–647.
3. Suzuki, N., Okayama, S., Nonaka, H., Tsuge, Y., Inui, M., and Yukawa, H. (2005) Large-scale engineering of the *Corynebacterium glutamicum* genome. *Appl. Environ. Microbiol.* **71**, 3369–3372.
4. Fukiya, S., Mizoguchi, H., and Mori, H. (2004) An improved method for deleting large regions of *Escherichia coli* K-12 chromosome using a combination of Cre/*loxP* and lambda Red. *FEMS Microbiol. Lett.* **234**, 325–331.
5. Yu, B. J., Sung, B. H., Koob, M. D., Lee, C. H., Lee, J. H., Lee, W. S., et al. (2002) Minimization of the *Escherichia coli* genome using a Tn*5*-targeted Cre/*loxP* excision system. *Nat. Biotechnol.* **20**, 1018–1023.
6. Goryshin, I. Y., Naumann, T. A., Apodaca, J., and Reznikoff, W. S. (2003) Chromosomal deletion formation system based on Tn*5* double transposition: use for making minimal genomes and essential gene analysis. *Genome Res*. **13**, 644–653.
7. Westers, H., Dorenbos, R., Dijl, J. M., Kabel, J., Flanagan, T., Devine, K. M., et al. (2003) Genome engineering reveals large dispensable regions in *Bacillus subtilis*. *Mol. Biol. Evol.* **20**, 2076–2090.
8. Lee, S. J., Lee, D. Y., Kim, T. Y., Kim, B. H., Lee, J., and Lee, S. Y. (2005) Metabolic engineering of *Escherichia coli* for enhanced production of succinic acid, based on genome comparison and *in silico* gene knockout simulation. *Appl. Environ. Microbiol*. **71**, 7880–7887.
9. Yoon, Y. G., Cho, J. H., and Kim, S. C. (1998) Cre/*loxP*-mediated excision and amplification of the *Escherichia coli* genome. *Genet*. *Anal.* **14**, 89–95.
10. Posfai, G., Koob, M., Hradecna, Z., Hasan, N., Filutowicz, M., and Szybalski, W. (1994) *In vivo* excision and amplification of large segments of the *Escherichia coli* genome. *Nucleic Acids Res.* **22**, 2392–2398.
11. Hasan, N., Koob, M., and Szybalski, W. (1994) *Escherichia coli* genome targeting, I. Cre-*lox*-mediated *in vitro* generation of *ori*-plasmids and their *in vivo* chromosomal integration and retrieval. *Gene* **150**, 51–56.
12. Datsenko, K. A., and Wanner, B. L. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6640–6645.
13. Goryshin, I. Y., Jendrisak, J., Hoffman, L. M., Meis, R., and Reznikoff, W. S. (2000) Insertional transposon mutagenesis by electroporation of released Tn*5* transposition complexes. *Nat. Biotechnol.* **18**, 97–100.
14. Miller, J. H. (1992) *A Short Course in Bacterial Genetics: A Laboratory Manual and Handbook for Escherichia coli and Related Bacteria*. Cold Spring Harbor, NY: Cold Spring HarborLaboratory Press.
15. Blattner, F. R., Plunkett, G. 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., et al. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474.
16. Hoffman, L. M., Jendrisak, J. J., Meis, R. J., Goryshin, I. Y., and Reznikoff, W. S. (2000) Transposome insertional mutagenesis and direct sequencing of microbial genomes. *Genetica* **108**, 19–24.
17. Armstrong, K. A. (1984) A 37 × 10^3 molecular weight plasmid-encoded protein is required for replication and copy number control in the plasmid pSC101 and its temperature-sensitive derivatives pHS1. *J. Mol. Biol*. **175**, 331–348.

18. Sambrook, J., and Russell, D. W. (2001) *Molecular Cloning: A Laboratory Manual*, 3rd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
19. Posfai, G., Kolisnychenko, V., Bereczki, Z., and Blattner F. R. (1999) Markerless gene replacement in *Escherichia coli* stimulated by a double-strand break in the chromosome. *Nucleic Acids Res*. **2**, 74409–74415.
20. Smalley, D. J., Whiteley, M., and Conway, T. (2003) In search of the minimal *Escherichia coli* genome. *Trends Microbiol*. **11**, 6–8.
21. Sung, B. H., Lee, C. H., Yu, B. J., Lee, J. H., Lee, J. Y., Kim, M. S., et al. (2006) Development of a biofilm production-deficient *Escherichia coli* strain as a host for biotechnological applications. *Appl. Environ. Microbiol.* **72**, 3336–3342.

18

Construction of Long Chromosomal Deletion Mutants of *Escherichia coli* and Minimization of the Genome

Jun-ichi Kato and Masayuki Hashimoto

Key Words: chromosome deletions; *Escherichia coli*; essential genes; FLP; recombinase; *red*.

1. Introduction

Genetic information consists of protein- and RNA-coding genes that exist in a range of sizes and noncoding *cis*- and *trans*-acting sequence elements. The use of long chromosomal deletion mutations is a powerful method for identifying essential genetic information through experimental reduction of the genome to its minimal gene set. Taking advantage of recent technical advances, we constructed sequence-specific long deletion mutations of the *Escherichia coli* chromosome. In a recent report (*1*), we described a set of *E. coli* medium-scale deletions (MDs) and large-scale deletions (LDs). Several LD mutations were combined to generate an engineered strain lacking ~30% of the parental chromosome. We then constructed another set of deletion mutations, MDs and small-scale deletions (SDs), and identified additional essential genetic regions using complementation analysis. To delete the remaining essential chromosomal regions, we developed an Flp recombinase target (FRT)-based system of site-specific recombination to move chromosomal regions onto mini-F plasmids *in vivo*. In this report, we describe the details of the construction of several of these types of large chromosomal deletion mutants.

2. Materials

1. *E. coli* strains: MG1655, MG1655 *rpsL polA12*, MG1655 *rpsL*, MG1655 *red* (*red*:*kan* (Δ(*recC ptr recB recD*)::*Plac-red*), MG1655 rsh3 (*red*:*tet* (Δ(*recC ptr recB recD*)::*Plac-red*) *rpsL hsdR*:Ap), and DH5α *pir*.

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

2. Plasmids: 664BSCK2-4, 415S Sm, mF-CRS, miniFtsFA, pFT-G, pSG76SA, miniFtsFAK, and 184Km *pir*.
3. Oligonucleotide primers.
4. Thermal cycler.
5. Ex Taq and LA Taq polymerases (TAKARA Shuzo, Kyoto, Japan).
6. Restriction enzymes and T4 DNA ligase.
7. Agarose gel equipment.
8. Media: LB broth and antibiotic medium 3 (Becton Dickinson, Sparks, MD).
9. Antibiotics: Ampicillin, chloramphenicol, kanamycin, streptomycin, tetracycline, and gentamicin.

3. Methods

We developed three systems for generating large deletion mutations of the *E. coli* chromosome based on the type of DNA used for transformation (plasmid or DNA fragments) and the recombination system used: (1) Plasmid system-1 used plasmid DNA and endogenous *E. coli* homologous recombination; (2) DNA fragment system used linear DNA fragments and λ phage homologous recombination (*red*); (3) Plasmid system-2 used plasmids and both endogenous *E. coli* homologous recombination and FLP recombinase-FRT (FLP-FRT) site-specific recombination. The methods below describe (a) the construction of the plasmids and/or DNA fragments used in these systems and (b) selection of recombinants.

3.1. Plasmid System-1

In Plasmid system-1, a deletion plasmid is constructed *in vitro*, containing sequences homologous to short regions flanking the desired chromosomal deletion. After introduction of the deletion plasmid into *E. coli*, recombinants are isolated, in which large regions of the parental chromosome are replaced by the corresponding sequences of the deletion plasmid (flanking sequences and intervening plasmid sequences, if any; **Fig. 1**). Replacement requires a two-step homologous recombination process: in the first step, recombinants, in which the deletion plasmid has integrated into the chromosome, are selected under conditions that prevented replication of the deletion plasmid; in the second step, recombinants, in which the integrated deletion plasmid has excised, are obtained using negative selection. Two Plasmid system-1 systems were developed, 664 (MD) and 415S Sm, that differed in their host strain and type of the plasmid replicon.

3.1.1. The 664 (MD) System

In this system, the host strain is a *polA rpsL* mutant, and the plasmid is a ColE1-replicon with its replication dependent on *polA*-encoded DNA polymerase I. Introduction of *rpsL* into an *E. coli polA* strain conferred sensitivity to streptomycin (the *rpsL*^{R} streptomycin-resistant, Sm^R mutant). The deletion plasmid, 664BSCK2-4 *(1)*, a derivative of pHSG664 *(2)*, contained two positive selection markers, chloramphenicol resistance (Cm^R) and kanamycin resistance (Km^R); two negative selection markers, *rpsL*$^+$ and *sacB* (**Note 1**); and two multiple-cloning sites flanking the Km^R marker (**Fig. 1A**).

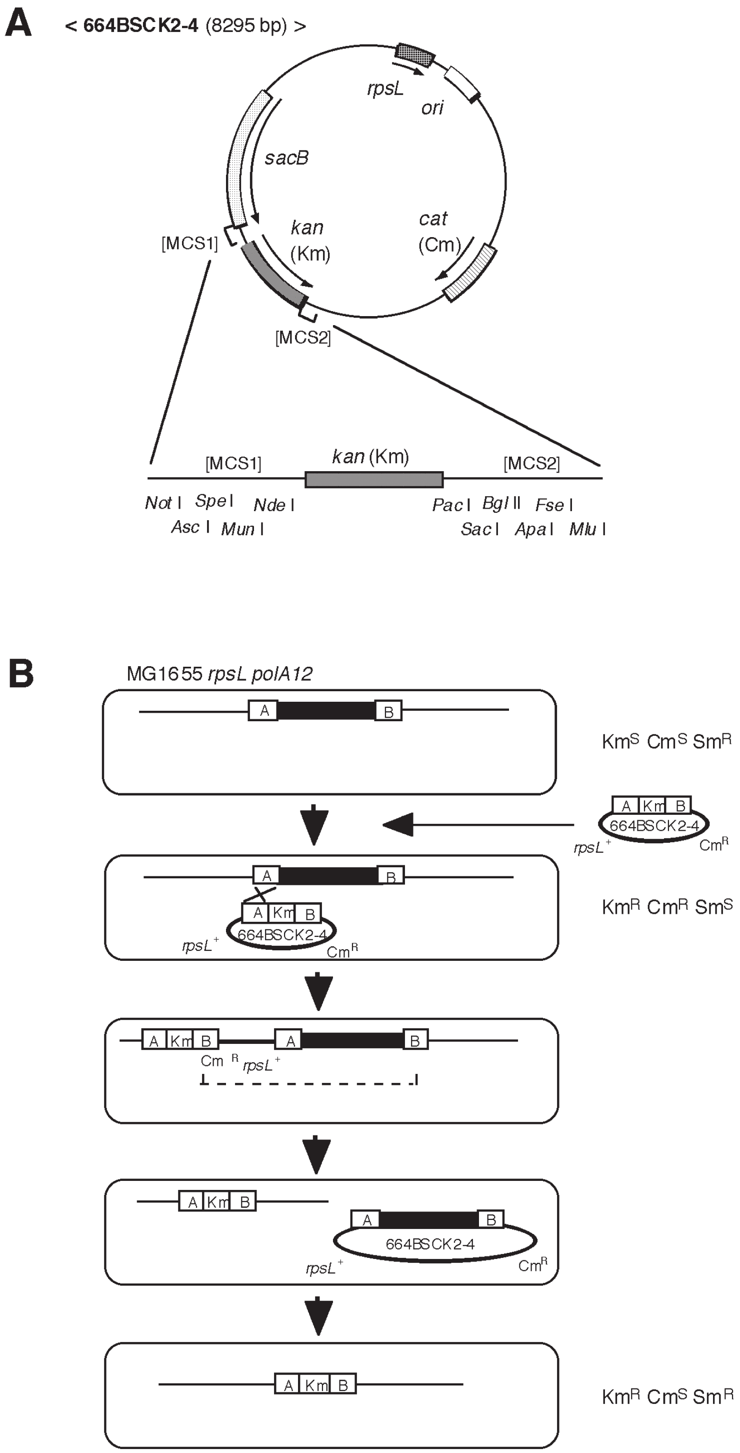


Fig. 1. The 664 (MD) system. **(A)** Schematic drawing of the 664BSCK2-4 plasmid *(1)*. This plasmid is a derivative of pHSG664, which has a ColE1 replicon and the *rpsL*$^+$ marker. *kan*, kanamycin-resistance gene (Km^R); *cat,* chloramphenicol-resistance gene (Cm^R). **(B)** Schematic of the process by which a deletion mutation is generated via homologous recombination using the 664BSCK2-4–based deletion plasmid. The chromosomal region to be deleted is represented by a bold line.

1. Nucleotide sequences flanking the chromosomal region to be deleted are amplified by polymerase chain reaction (PCR) using primers containing appropriate restriction enzyme sites, and these sequences are subcloned into 664BSCK2-4.
2. Recombinant plasmids are introduced into *E. coli* strain MG1655 *rpsL polA12* (**Note 2**).
3. Cm^R transformants are selected at 42°C (**Note 3**).
4. Transformants are incubated at 30°C to obtain Km^R Cm^S Sm^R recombinants (**Fig. 1B**).

3.1.2. The 415S Sm System

A temperature sensitive (ts)-replication plasmid is used in this system, allowing the use of a wide range of host strains. Plasmid 415S Sm was constructed by cloning wild-type *rpsL* allele into the ts plasmid pHSG415S (**Fig. 2A**).

1. The chromosomal regions flanking the region to be deleted are inserted into the *Nru*I site of 415S Sm (**Note 4**). By joining the two flanking regions directly and then inserting them into the *Nru*I site, a markerless chromosomal deletion plasmid is created (**Fig. 2**).
2. The deletion plasmid is introduced into *E. coli* strain MG1655 *rpsL*.
3. Cm^R transformants are selected at 42°C (**Note 3**).
4. Transformants are then incubated at 30°C to obtain the Sm^R derivatives (**Fig. 2B**).

Note, that it was particularly important to confirm the genomic structure of the introduced deletion by PCR when using the markerless deletion construct.

3.2. DNA Fragment System

Wild-type *E. coli* has an active DNA exonuclease, Exo V of the RecBCD complex, which quickly degrades any linear DNA that is introduced into the cell. Degradation of linear DNA can be circumvented in Exo V mutants, but the frequency of homologous recombination is very low in these strains, as Exo V is required for recombination. Introduction of an artificial homologous recombination system into Exo V mutant strains makes them suitable for chromosome engineering using linear DNA fragments. The λ phage homologous recombination system (*red*) has been shown to enhance recombination frequency of Exo V mutants and has the added feature of being able to mediate recombination between very short regions of homology.

3.2.1. SD System

The SD system we use is a one-step gene replacement strategy using PCR-generated targeting constructs and has been described previously ***(3–5)***.

1. A linear DNA fragment encoding the Cm^R gene is generated by PCR using oligonucleotide primers consisting of a 40-base-pair region of homology to the flanking regions of the deletion target site and 20 additional nonhomologous base pairs at the 3′ terminus of the primer (**Note 5**).
2. Linear DNA fragments are introduced into *E. coli* strain MG1655 *red* by electroporation, and Cm^R recombinants are isolated (**Note 6**).

MG1655 *red* is a derivative of strain KM22 in which the expression of the *red* genes (α and β) is under the regulation of the *lac* promoter ***(6)***. Isopropylthio-β-D-galactoside (IPTG; 25 to 100 μg/mL) is added to the culture media before and after electroporation. In this system, when a chromosomal region containing an essential gene is successfully

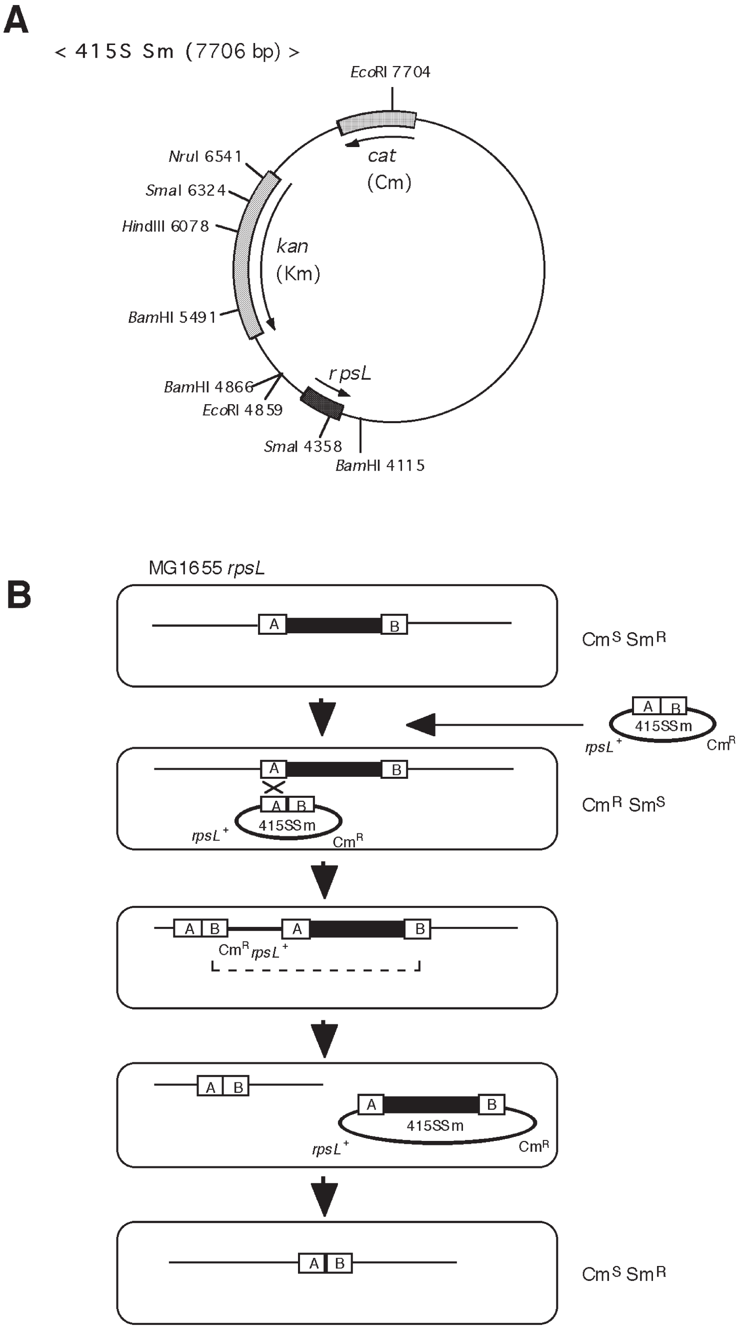


Fig. 2. The 415S Sm system. (**A**) Schematic drawing of the 415S Sm plasmid. This plasmid is a derivative of pHSG415S, which has a temperature-sensitive pSC101 replicon and the *rpsL*$^+$ marker. *kan*, Km^R; *cat*, Cm^R. (**B**) Schematic of the process by which a deletion mutation is generated using the plasmid 415S Sm–based deletion construct. The chromosomal region to be deleted is represented by a bold line.

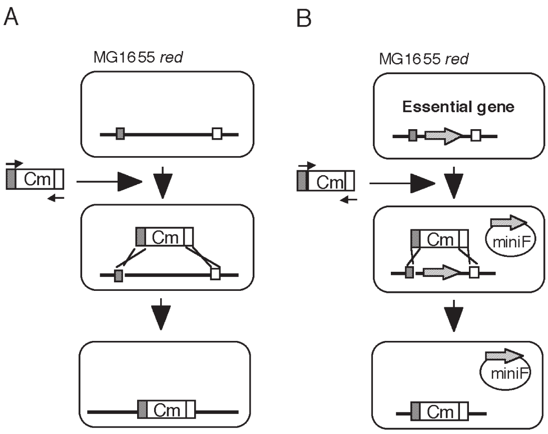


Fig. 3. SD system. Schematic of the construction of a deletion mutation in the absence **(A)** or presence **(B)** of a complementing plasmid for the essential gene.

targeted, the essential gene is excised into a mini-F plasmid, and normal function is maintained in the presence of this complementing plasmid (**Fig. 3B**).

3.2.2. LD System (CRS Cassette Method)

The LD system, which has been used to reduce the overall size of the genome ***(1)***, utilizes markerless deletion constructs (**Fig. 4**). Markerless deletion mutations are useful when targeting multiple chromosomal regions to reduce the size of the genome, because single-marker genes can be used only for one round of selection and because many copies of a marker gene on the chromosome could lead to undesirable secondary DNA rearrangements. In the LD system, deletion units are generated using PCR and then combined to yield a single deletion construct.

1. In **step 1**, two rounds of PCR are carried out to generate a DNA fragment, in which the chromosomal sequences flanking the region to be deleted are joined to the sides of a Cm^R-*rpsL-sacB* (CRS) cassette (**Fig. 5A**). The CRS cassette is approximately 5 kb in size and carries a positive selection marker, Cm^R, and two negative selection markers, *rpsL*$^+$ and *sacB*. Introduction of *rpsL*$^+$ and *sacB* into a host strain confers sensitivity to streptomycin and sucrose, respectively ***(1)***.
2. The linear deletion construct is introduced into strain MG1655 rsh3 by electroporation.
3. Cm^R colonies are selected, and the structure of the deleted region of the chromosome is confirmed by PCR.
4. In **step 2**, two rounds of PCR are carried out to join two homologous flanking chromosomal regions directly (**Fig. 5B**).
5. This cassetteless linear DNA construct is introduced into the Cm^R strains obtained during **step 1**, and Sm^R and sucrose-resistant colonies are selected (**Note 7**).
6. The structure of the deleted region is again confirmed by PCR. For each deletion construct, the final genomic structure of the deleted region of the chromosome is different. In **step 1**, a chromosomal region is replaced with a CRS cassette, and in **step 2**, it is not.
7. To combine deletions, P1 phage is propagated on strains with or without the CRS cassette. Chromosomal deletions, in which the deleted regions have been replaced by the CRS

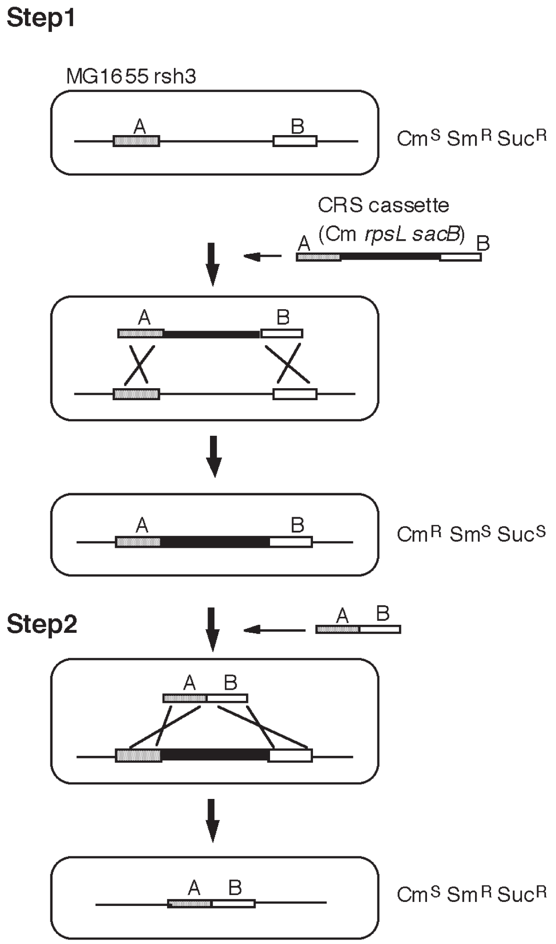


Fig. 4. LD system. Schematic of method for generating deletion units using the CRS cassette. The chromosomal region to be deleted is represented by a bold line.

cassette, are introduced by transduction with P1 phage prepared from CRS cassette-containing deletion mutants and selected using the positive selection marker Cm^R.

8. This is followed by transduction with P1 phage prepared from cassetteless deletion mutants, and recombinants, in which the Cm^R marker has been lost, are isolated using the negative selection markers *rpsL*$^+$ (Sm^R) and *sacB* (sucrose resistance).

3.3. Plasmid System-2

If recombinants containing a chromosomal deletion were obtained, then the deleted region did not contain an essential gene. On the other hand, the fact that a deletion mutation is nonviable does not provide a great deal of information about the structure or identity of the essential genetic element(s) contained in that particular region of the chromosome. For example, we cannot determine whether there was a *trans*-acting or *cis*-acting element in that region. However, in the case of essential *cis*-acting genetic elements, such as the origin of replication (*oriC*), if a chromosomal region can be

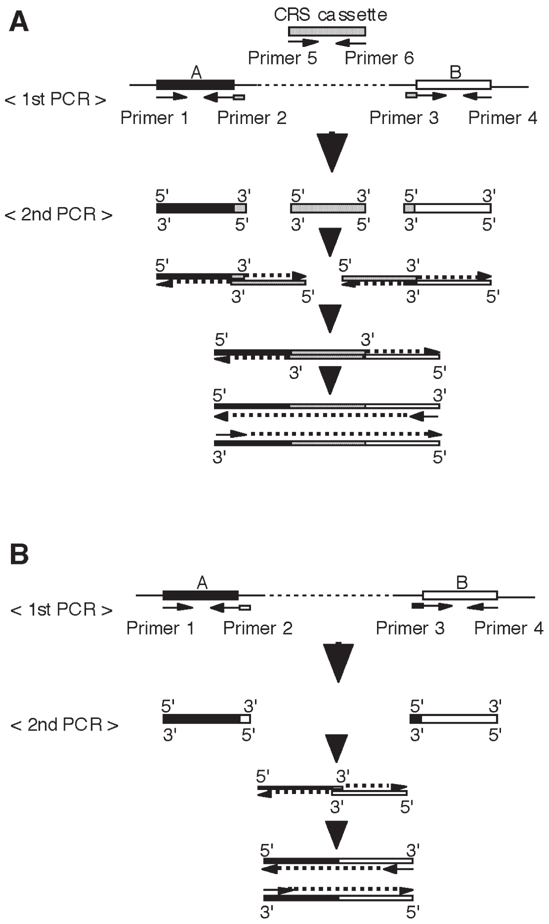


Fig. 5. Schematic of the generation of the CRS cassette–containing **(A)** and the simple **(B)** deletion constructs using PCR. Boxes *A* and *B* represent chromosomal regions of homology, and the gray box represents the CRS cassette. The arrows indicate the position of the primers used to amplify each fragment.

excised and transferred to an extrachromosomal plasmid *in vivo*, we can reasonably conclude that there are no essential *cis*-acting elements in that region. We generated a system for testing the presence of essential *cis*-acting elements, which is similar to other methods used for cloning and/or deletion analysis *(7–11)*. The unique feature of our system is the use of mini-F plasmids. Because we can maintain the mini-F plasmid as a single-copy plasmid, we were able to analyze the excised region in its near-native state.

3.3.1. FRT1 System

The FRT1 system, our prototype FRT system, uses three different plasmids (**Fig. 6A**). The mini-F plasmid, miniFtsFA, was constructed by inserting the *Hin*dIII fragment of pSG76-A *(8)*, containing *bla* and an FRT site, into a ts mini-F plasmid, 7577, which is replication defective at 42°C. The structure of mini-F plasmid 7577 is identical to pKP1592, except for the unidentified ts mutation ***(12)***. Plasmid pSG76SA, which is related to pR6K, was constructed by ligating the *Apa*LI-*Cla*I fragment of pSG76-A, containing an incomplete copy of *bla* and an FRT site, to the *Apa*LI-*Nar*I fragment of pIB279, containing *sacB*. The resultant plasmid, pSG76SA, does not encode *pir*, which is necessary for replication. Replication is possible only if the Rep protein, which is encoded by *pir*, is supplied. Plasmid pFT-G, which is related to pSC101, was obtained by inserting the *Pvu*II-*Sma*I fragment of pAB2001, containing the gentamicin-resistance gene (*aacC1*), into the *Sca*I site of pFT-A ***(8)***, which contains the gene for FLP recombinase. This plasmid is also temperature sensitive and is replication deficient at temperatures >34°C.

In the FRT-1 system, the two chromosomal regions (A and B in **Fig. 6B** and **Fig. 7**) flanking the region to be deleted are ligated onto either end of the kanamycin-

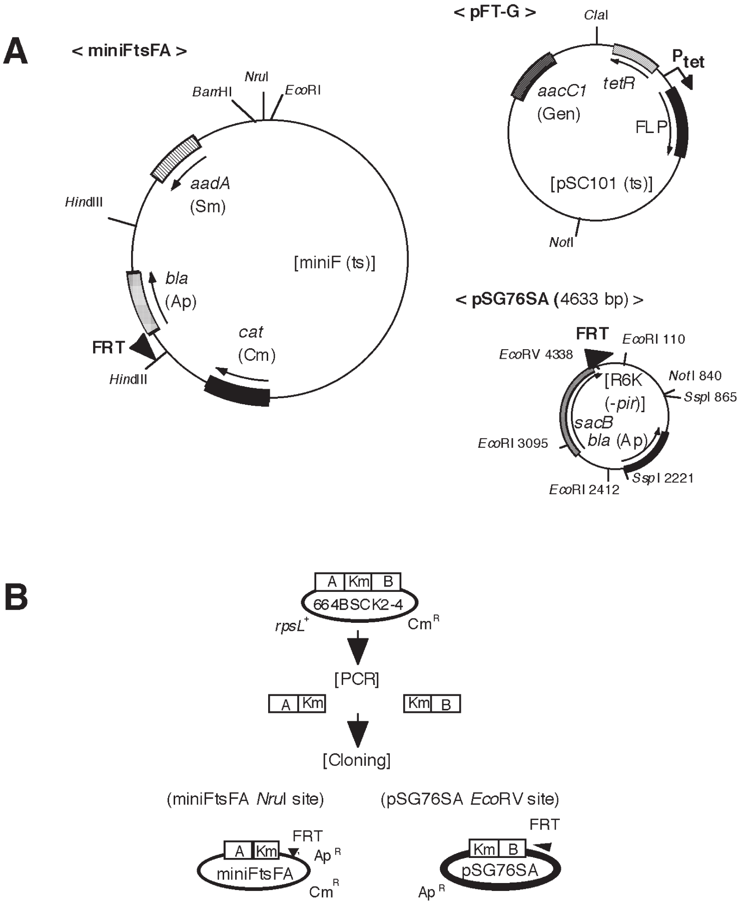


Fig. 6. Plasmids for the FRT1 system. **(A)** Schematic drawing of the three plasmids used in the FRT system: miniFtsFA, pFT-G, and pSG76SA. *bla*, Ap^R; *aadA*, Sm^R; *cat*, Cm^R; *aacC1*, Gen^R. **(B)** Schematic of the process of cloning the two flanking chromosomal regions (**A** and **B** in **Fig. 7**) into their respective deletion plasmids.

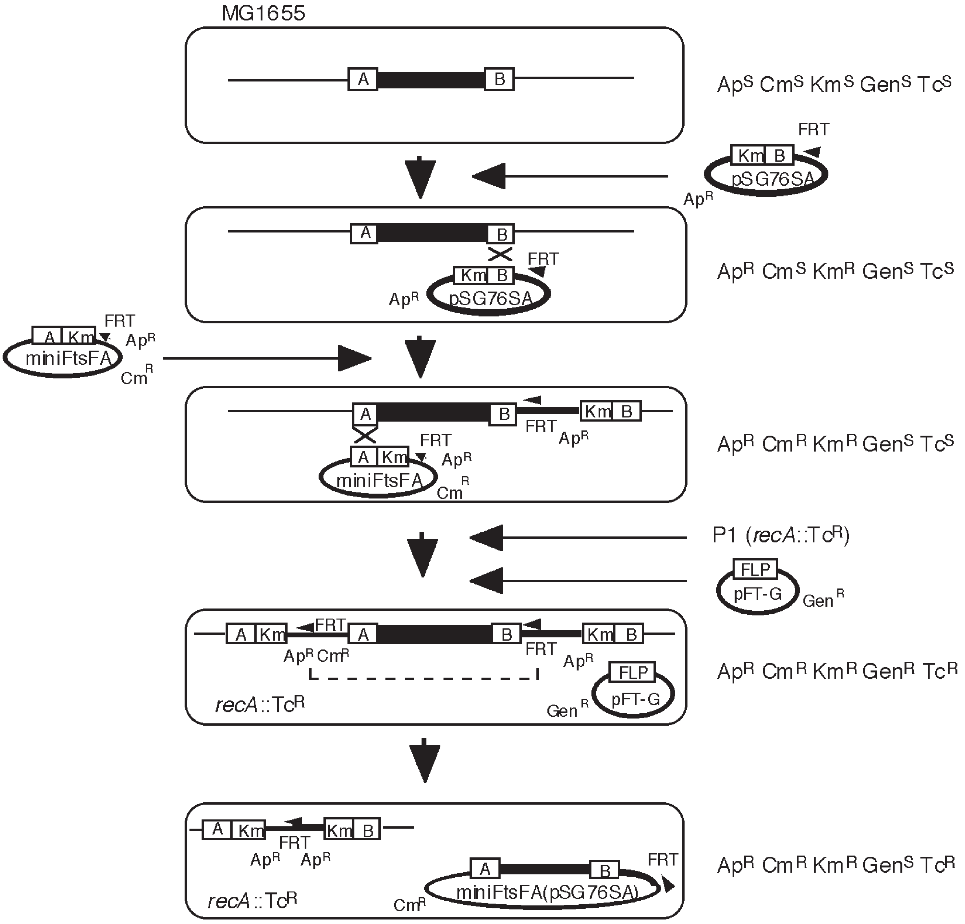

MG1655
A
B
Ap^S Cm^S Km^S Gen^S Tc^S
FRT
Km B
pSG76SA
Ap^R
Ap^R Cm^S Km^R Gen^S Tc^S
A Km
miniFtsFA
Cm^R
Ap^R Cm^R Km^R Gen^S Tc^S
P1 (*recA*::Tc^R)
FLP
pFT-G
Gen^R
recA::Tc^R
Ap^R Cm^R Km^R Gen^R Tc^R
miniFtsFA(pSG76SA)
Ap^R Cm^R Km^R Gen^S Tc^R

Fig. 7. FRT system. Schematic of the transfer of a chromosomal region to a mini-F plasmid. The chromosomal region to be deleted is represented by the bold line.

resistance gene to create A-Km and B-Km, using PCR and the p664BSCK2-4-derivative as the template (Fig. 1 and Fig. 6B). A-Km is inserted into the *Nru*I site of miniFtsFA, and B-Km was inserted into the *Eco*RV site of pSG76SA (Fig. 6B).

1. In **step 1**, pSG76SA, carrying B-Km, is introduced into wild-type strain MG1655. The plasmid cannot replicate in this strain due to lack of *pir* (Fig. 7). Km^R recombinants, in which the plasmid has integrated into the chromosome by homologous recombination, are isolated.
2. Next, miniFtsFA, carrying A-Km, is introduced into Km^R-resistant colonies (obtained in **step 1**), and Cm^R colonies, representing second step recombinants, are obtained after incubation at 42°C (Fig. 7).
3. To inhibit homologous recombination beyond this stage, *recA* is disrupted by P1 transduction.
4. The FLP-containing plasmid is then introduced into the second step recombinants (Fig. 7).

5. Addition of tetracycline to the culture media results in expression of FLP recombinase and simultaneous plasmid excision and a chromosomal deletion.

3.3.2. FRT2 System

The FRT2 system is similar to FRT1 but with some improvements. The mini-F plasmid miniFtsFAK was constructed by converting Cm^R to Km^R (derived from miniFtsFA) using *red* recombination (**Fig. 8A**). In the FRT2 system, the two chromosomal regions (A and B in **Fig. 8** and **Fig. 9**) flanking the region to be deleted are generated by PCR.

1. Two or three rounds of PCR are carried out to prepare (76 arm)-A-Cm-(mF arm) and B-Cm (**Fig. 8B**), respectively.

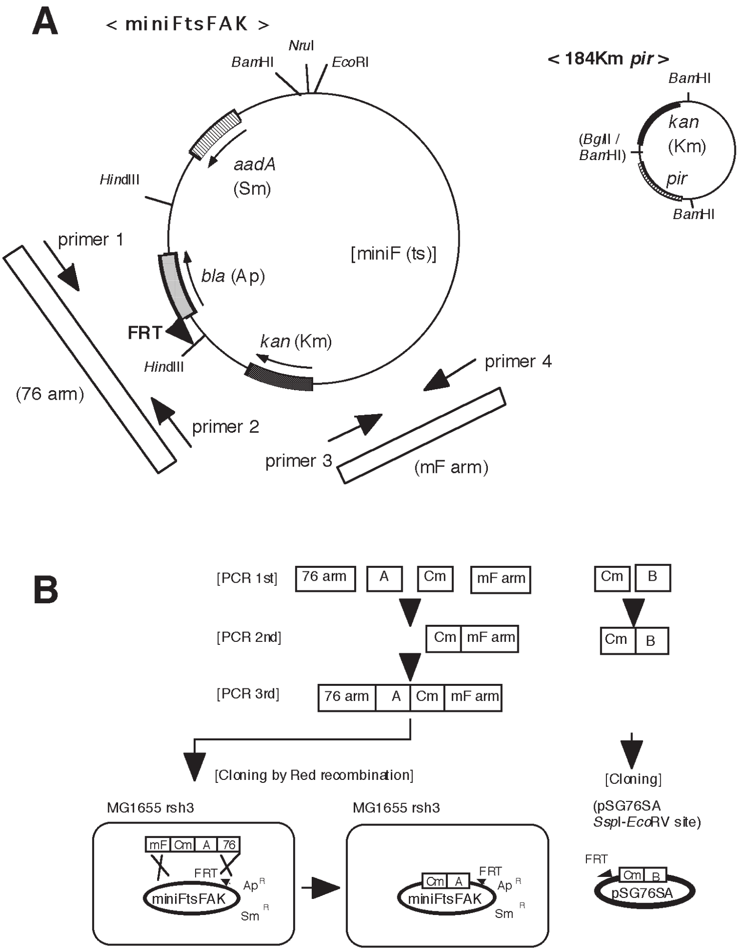


Fig. 8. Plasmids for FRT2 system. **(A)** Schematic drawing of the plasmids used in the FRT2 system: miniFts-FAK, and p184 Km *pir*. *bla*, Ap^R; *aadA*, Sm^R; *kan*, Km^R. **(B)** Schematic of the cloning of the two flanking chromosomal regions (**A** and **B** in **Fig. 9** and **Fig. 10**) into their respective deletion plasmids.

2. A-Cm is inserted into miniFtsFAK by *red* homologous recombination.
3. B-Cm is inserted into pSG76SA by ligating it with the *Ssp*I-*Eco*RV fragment of pSG76SA (**Fig. 8B**).
4. In **step 1** (**Fig. 9**), pSG76SA carrying B-Cm is introduced into wild-type strain MG1655. Similar to the FRT1 system, plasmid replication in this strain is repressed due to the absence of *pir*. Cm^{R} colonies (**step 1** recombinants) are isolated, in which the plasmid has integrated into the chromosome by homologous recombination.
5. Next, miniFtsFAK carrying A-Cm is introduced into **step 1** recombinants, and ampicillin-resistant (Ap^{R}) transformants are obtained at 30°C.
6. Ap^{R} Sm^{R} colonies, representing **step 2** recombinants, are then isolated at 42°C (**Fig. 9**).
7. To inhibit homologous recombination beyond this stage, *recA* is disrupted by P1 transduction.

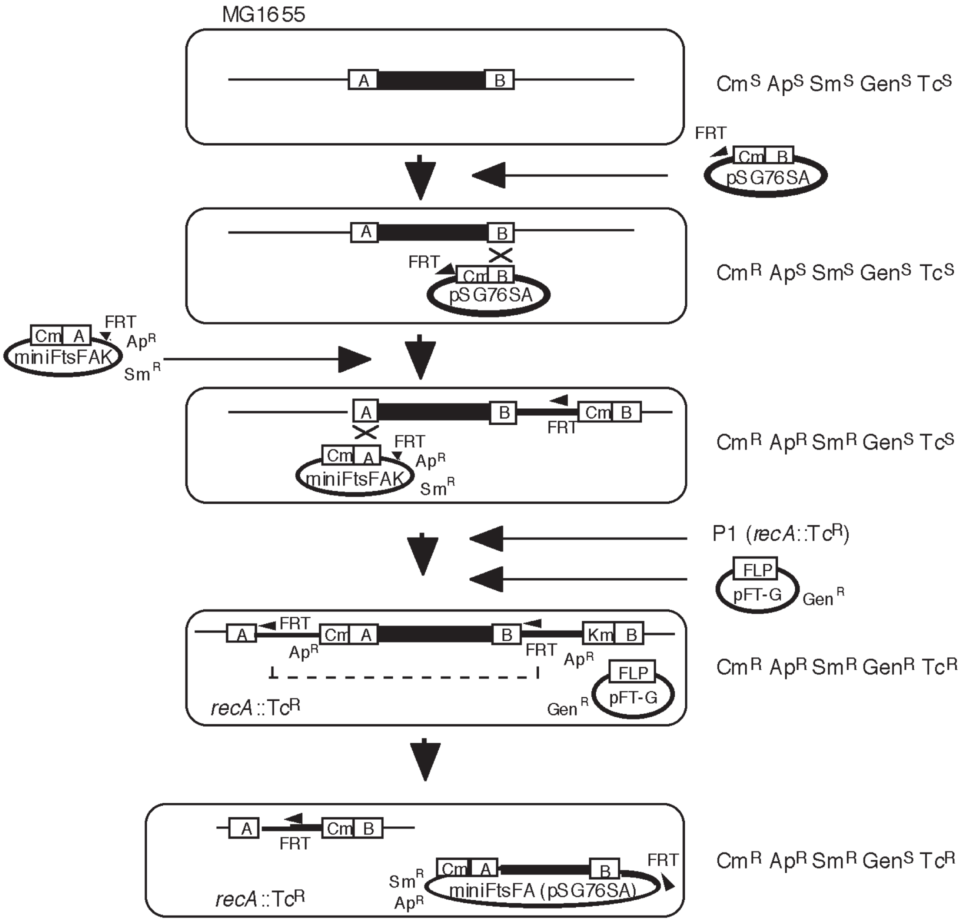


Fig. 9. FRT2 system. Schematic of the transfer of a chromosomal region to a mini-F plasmid. The chromosomal region to be deleted is represented by the bold line.

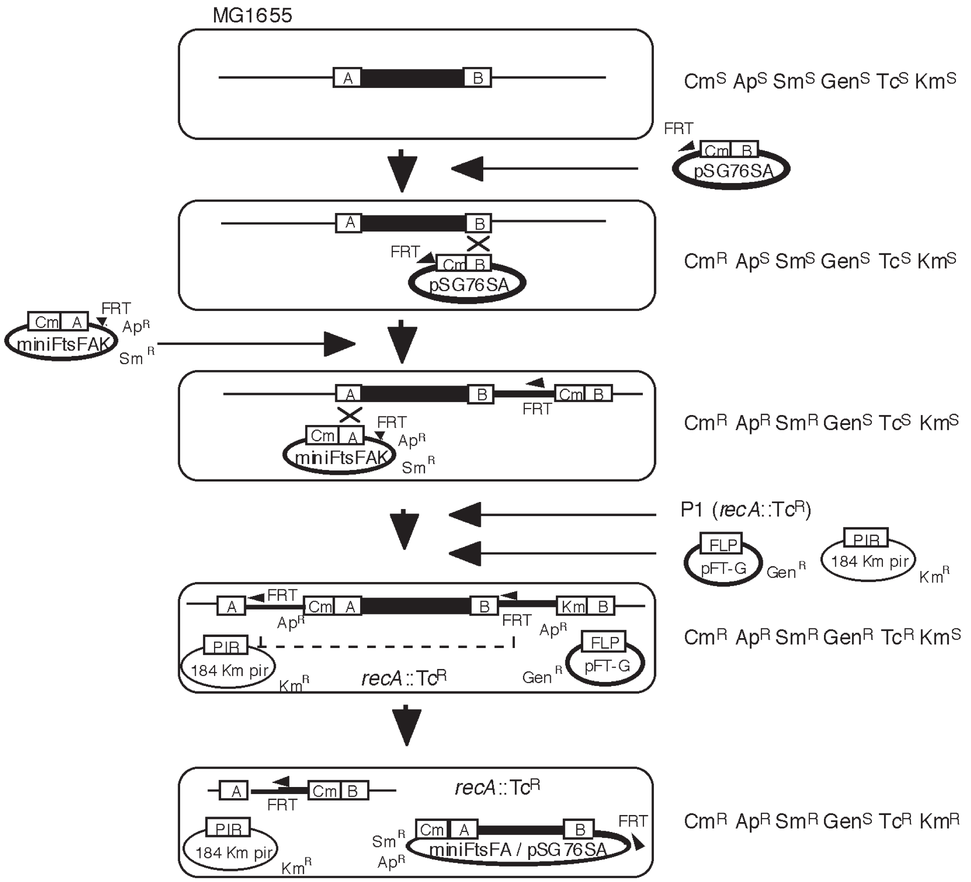

MG1655
A
B
Cm^S Ap^S Sm^S Gen^S Tc^S Km^S
FRT
Cm B
pSG76SA
Cm^R Ap^S Sm^S Gen^S Tc^S Km^S
Cm A
FRT
Ap^R
Sm^R
miniFtsFAK
Cm^R Ap^R Sm^R Gen^S Tc^S Km^S
P1 (*recA*::Tc^R)
FLP
pFT-G
Gen^R
PIR
184 Km pir
Km^R
Km B
recA::Tc^R
Cm^R Ap^R Sm^R Gen^R Tc^R Km^S
miniFtsFA / pSG76SA
Cm^R Ap^R Sm^R Gen^S Tc^R Km^R

Fig. 10. FRT3 system. Schematic of the transfer of a chromosomal region to a mini-F plasmid. The chromosomal region to be deleted is represented by the bold line.

8. The FLP-plasmid is introduced into **step 2** recombinants (**Fig. 9**), and tetracycline is added to the culture media to induce expression of the FLP recombinase and simultaneous plasmid excision and chromosome deletion (**Fig. 9**).
9. Finally, to obtain a strain free of pFT-G, cells are incubated at 35°C, at which point pFT-G does not replicate, but the mini-F ts replicon remains functional.

3.3.3. FRT3 System

In the FRT2 system, the excised plasmid can replicate as a miniF(ts) replicon because the pSG76SA-derived R6K replicon is nonfunctional due to lack of *pir* in the host strain. For a few of the chromosomal deletions, the excised plasmid was not maintained well and the resultant chromosome deletion was not obtained. To address this issue, in the FRT3 system, p184 Km *pir*, encoding a functional copy of *pir*, was cointroduced with pFT-G (**Fig. 10**), and the excised plasmid containing the R6K replicon was maintained. In all other aspects, the FRT3 system was the same as the FRT2 system.

Notes

1. The *sacB* gene on p664BSCK2-4 is not fully functional. It does not confer a clear host sensitivity to sucrose.
2. The *polA12* is a temperature-sensitive mutation. However, ColE1-related plasmids replicate poorly even at 30°C.
3. The isolation frequency of recombinants at this step is dependent on the method used to introduce the plasmids into bacteria. At high transformation efficiencies (e.g., with electroporation), recombinants can be obtained directly after incubation at 42°C. If transformation efficiencies are low, however, transformants need to be first selected at 30°C, followed by selection of recombinants at 42°C.
4. Cloning of chromosomal fragments into the 415S Sm is not successful 100% of the time. Although the underlying reason requires additional analysis, it appears that certain DNA fragments are refractory to subcloning into a pSC101 replicon.
5. The frequency of recombination was low using primers containing a 40-base-pair region of homology but improved upon attachment of a ~1-kb region of homology to either end of the Cm^R gene.
6. In the case of Cm selection, the Cm^R recombinants often appeared after a few days of incubation at room temperature. Antibiotic medium 3 was more useful for long-term incubation.
7. The medium for selection of Sm^R and sucrose-resistant colonies was LB containing 10% sucrose and streptomycin (50 µg/mL), and no NaCl.

Acknowledgments

The authors thank Dr. Takeyoshi Miki for providing the original mini-F ts plasmid (plasmid number is 7577). We also thank for Dr. Posfai for providing the strain DH5α *pir* and the plasmids pSG76-A, pFT-A, and pPIR-A. This work was supported by KAKENHI a (Grant-in-Aid for Scientific Research) on Priority Areas "Systems Genomics" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Hashimoto, M., Ichimura, T., Mizoguchi, H., Tanaka, K., Fujimitsu, K., Keyamura, K., et al. (2005) Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. *Mol. Microbiol.* **55**, 137–149.
2. Hashimoto-Gotoh, T., Kume, A., Masahashi, W., Takeshita, S., and Fukuda, A. (1986) Improved vector, pHSG664, for direct streptomycin-resistance selection: cDNA cloning with G:C-tailing procedure and subcloning of double-digest DNA fragments. *Gene* **41**, 125–128.
3. Datsenko, K. A., and Wanner, B. L. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6640–6645.
4. Murphy, K. C., Campellone, K. G., and Poteete, A. R. (2000) PCR-mediated gene replacement in *Escherichia coli*. *Gene* **246**, 321–330.
5. Yu, D., Ellis, H. M., Lee, E. C., Jenkins, N. A., Copeland, N. G., and Court, D. L. (2000) An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5978–5983.

6. Murphy, K. C. (1998) Use of bacteriophage λ recombination functions to promote gene replacement in *Escherichia coli*. *J. Bacteriol.* **180**, 2063–2071.
7. Ayres, E. K., Thomson, V. J., Merino, G., Balderes, D., and Figurski, D. H. (1993) Precise deletions in large bacterial genomes by vector-mediated excision (VEX). The *trfA* gene of promiscuous plasmid RK2 is essential for replication in several gram-negative hosts. *J. Mol. Biol.* **230**, 174–185.
8. Posfai, G., Koob, M., Hradecna, Z., Hasan, N., Filutowicz, M., and Szybalski, W. (1994) *In vivo* excision and amplification of large segments of the *Escherichia coli* genome. *Nucleic Acids Res.* **22**, 2392–2398.
9. Wild, J., Hradecna, Z., Posfai, G., and Szybalski, W. (1996) A broad-host-range *in vivo* pop-out and amplification system for generating large quantities of 50- to 100-kb genomic fragments for direct DNA sequencing. *Gene* **179**, 181–188.
10. Posfai, G., Koob, M. D., Kirkpatrick, H. A., Blattner, F. R. (1997) Versatile insertion plasmids for targeted genome manipulations in bacteria: isolation, deletion, and rescue of the pathogenicity island LEE of the *Escherichia coli* O157:H7 genome. *J. Bacteriol.* **179**, 4426–4428.
11. Yoon, Y. G., Cho, J. H., and Kim, S. C. (1998) Cre/*loxP*-mediated excision and amplification of large segments of the *Escherichia coli* genome. *Genet. Anal.* **14**, 89–95.
12. Murayama, N., Shimizu, H., Takiguchi, S., Baba, Y., Amino, H., Horiuchi, T., et al. (1996) Evidence for involvement of *Escherichia coli* genes *pmbA, csrA* and a previously unrecognized gene *tldD*, in the control of DNA gyrase by *letD* (*ccdB*) of sex factor F. *J. Mol. Biol.* **256**, 483–502.

19

Identification of Essential Genes in *Staphylococcus aureus* by Construction and Screening of Conditional Mutant Library

Dezhong Yin and Yinduo Ji

Summary

Antisense RNA technology has been used effectively to downregulate gene expression in a variety of bacterial systems. Regulated antisense RNA strategy provides an important approach to identify and characterize essential genes critical to bacterial growth *in vitro* and *in vivo*. This strategy allows selective genes to be turned on or off and to be expressed at certain levels. The availability of the *Staphylococcus aureus* (*S. aureus*) genome sequence makes it feasible to generate a gene-specific antisense RNA library. The combination of regulated antisense RNA technology and the gene-specific antisense RNA library allows for genome-wide analyses of functions of staphylococcal gene products for growth in culture and survival during infection.

Key Words: antisense RNA; essential genes; genome; *Staphylococcus aureus*.

1. Introduction

Various strategies involving knockout methods have been developed to inactivate gene products and to evaluate the importance of genes for bacterial viability *in vitro* (*1–4*). However, mutations in genes essential for growth are not viable for further evaluation of potential antibiotic drug targets. Therefore, conditional disruption of gene expression by using regulated antisense RNA is an important approach for addressing information on genes essential for bacterial growth or pathogenesis. A regulated antisense RNA expression system, which places genes downstream of the xylose/tetracycline chimeric promoter in an antisense orientation, allows selective genes to be turned off and to be expressed at certain levels to provide quantitative data on the gene product (*5, 6*). The availability of the *Staphylococcus aureus* genome sequence makes it feasible to identify all essential genes in *S. aureus* by screening a genome-wide library of antisense RNA-expressing strains (*[7–11]*; *see* **Chapter 20**). In this chapter, we would like to use the construction of enoyl–acyl carrier protein reductase (*fabI*, a key enzyme in the essential fatty acid biosynthesis pathway) antisense RNA as an example to identify essential genes in *S. aureus*.

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

2. Materials

1. Luria-Britani (LB) medium (BD Biosciences; Sparks, MD).
2. Tryptic soy broth (TSB) medium (BD Biosciences).
3. Erythromycin (Erm).
4. Tetracycline (Tc).
5. Anhydrotetracycline (ATc).
6. *E. coli* ElectroMax DH10B (Invitrogen, Carlsbad, CA).
7. pYH4 (an *E. coli*/*S. aureus* shuttle vector) ***(12)***.
8. RN4220/tetA (a laboratory *S. aureus* strain accepting foreign DNA directly from *E. coli*) ***(12)***.
9. WCUH29 (a clinical *S. aureus* strain) ***(13)***.
10. Oligonucleotide primers.
11. High-fidelity *pfx* DNA polymerase and T4 DNA ligase (Invitrogen).
12. *Asc* I and *pme* I restriction enzymes (New England BioLabs, Beverly, MA).
13. Agarose gel and DNA sequencing equipment.
14. PCR machine (MJ Research, Waltham, MA).
15. Gene Pulser (Bio-Rad, Hercules, CA).
16. QIAprep Miniprep Kit (Qiagen, Valencia, CA).
17. PCR purification kit (Qiagen).
18. TSB-Erm agar (TSA-Erm) plates.
19. LB-Amp agar plates.
20. Microtiter plate reader SpectraMax plus384 (Molecular Devices, Sunnyvale, CA).
21. PCR Supermix (Invitrogen).
22. SOC medium (Invitrogen).
23. Phosphate-buffered saline (PBS; Sigma, St. Louis, MO).
24. CD1-female mice (25 g) (Charles River Laboratories, Wilmington, MA).

3. Methods

The methods described below outline (1) construction of regulated antisense RNA expression plasmid, (2) expression of antisense RNA, and (3) characterization of antisense RNA mutants.

3.1. Construction of S. aureus fabI *Antisense Expressing Vectors*

3.1.1. Primer Design for Antisense Fragments

The *S. aureus* and other microbial genomes have been completed and are available in a public computer domain: (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genome&cmd=Retrieve&dopt=Overview&list_uids=179). It is possible to construct a gene-specific antisense expression vector or a gene-specific antisense expression library based on open reading frame (ORF) information at the NCBI genome bank. It is clear that the efficacy of antisense RNA inhibition depends on the region of ORF and length of antisense RNA fragments ***(12)***. To compare the effects of expression of different *fabI* antisense RNA fragments on *S. aureus* phenotype, we designed three pairs of PCR primers for three different antisense fragments as follows (**Note 1**): (I) *fabI*for1 (5′ GTT**GGCGCGCC**GGGATTAGATATTCTATCCG 3′; 51 bp upstream of the *fabI* start codon; the boldface sequence in the oligonucleotide primers is *Asc* I recognition sequence) and *fabI*rev1 (5′GAGCCAC AATTGTTAATGAG 3′; 389 bp downstream of the start codon of *fabI*); (II) *fabI*for2 (5′GGTT**GGCGCGCC**ATATGTCATCATGGGAATCG 3′; 24 bp downstream of start codon of *fabI*) and *fabIrev1*; (III) *fabI*for3 (5′GGTT**GGCGCGCC**GTTCAAAGCGATGAAGAGG 3′; 209 bp downstream of *fabI* start codon) and *fabI*rev2 (5′GCGTGGAATCCGCTATCTACATG 3′; 14 bp upstream of *fabI* stop codon).

3.1.2. PCR Synthesis of DNA Fragments

These pairs of primers are used to amplify DNA fragments, including three different *fabI* fragments, by using *S. aureus* RN4220 genomic DNA as a template, high-fidelity *pfx* DNA polymerase (Invitrogen), and a PCR thermocycler (MJ Research) (**Note 2**).

3.1.3. Clone of DNA Fragments into Tc-Inducible Vector, pYH4, in Antisense Orientation

1. Purification of PCR products using PCR purification kits (Qiagen) per the manufacturer's instruction.
2. Digestion of the purified PCR products with the *Asc* I restriction enzyme (New England BioLabs) per the manufacturer's instruction.
3. Purification of *Asc* I–digested PCR products by using a PCR purification kit (Qiagen).
4. Ligation of the *Asc* I–digested PCR products and the *Asc I/Pme* I–digested Tc-inducible vector, pYH4 *(12)*, a plasmid carrying Erm resistance marker and able to replicate in *E. coli* and *S. aureus* (**Fig. 1** and **Note 3**).
5. Electroporation of the ligated DNA into 25 µL of *E. coli* ElectroMax DH10B cells in a 0.1-cm cuvette at 1.8 kV, 200 Ω, and 25 µF using the Gene Pulser unit (Bio-Rad).
6. Transformed cells are incubated with 900 µL of SOC medium at 37°C for 45 min and plated (25 µL) onto LB-agar plates (Erm, 300 µg/mL). The plates are incubated for 24 to 48 h.

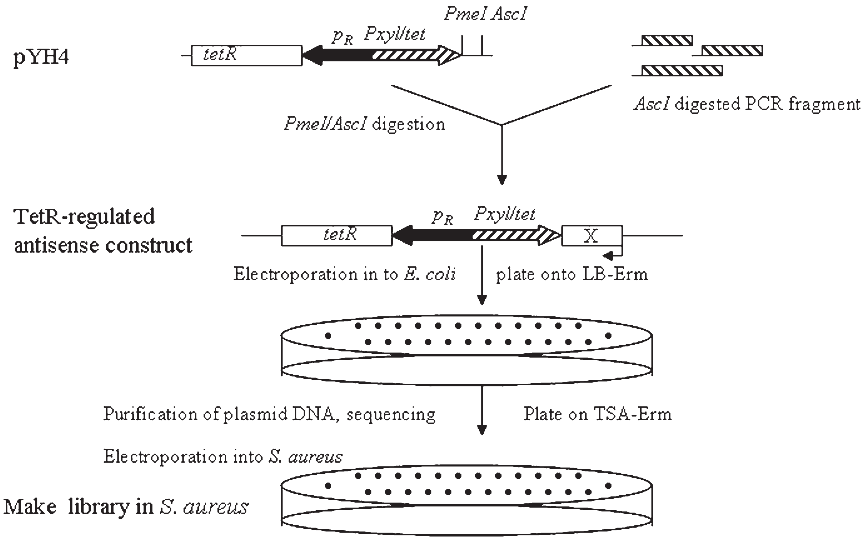

PmeI AscI
pYH4
tetR
pR Pxyl/tet
AscI digested PCR fragment
PmeI/AscI digestion
TetR-regulated antisense construct
tetR
pR Pxyl/tet
X
Electroporation in to E. coli
plate onto LB-Erm
Purification of plasmid DNA, sequencing
Plate on TSA-Erm
Electroporation into S. aureus
Make library in S. aureus

Fig. 1. Construction of *S. aureus* antisense expression library. Different gene fragments are generated by PCR, digested with *Asc* I, and ligated into the *Asc* I and *Pme* I sites of pYH4 in an antisense orientation. (Adapted from Ref. *12* by permission of Blackwell Publishing.)

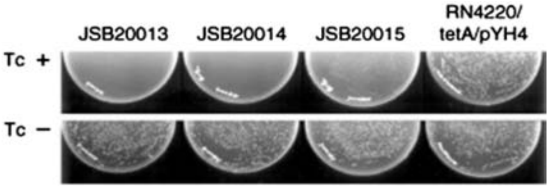


Fig. 2. Screening the *S. aureus fabI* antisense strains for conditional lethal phenotypes. Overnight cultures of *S. aureus* strains were diluted and plated onto TSA-Erm plates in the presence or absence of inducer Tc (1 µg/mL) and incubated at 37°C overnight. (Reprinted from Ref. ***12*** by permission of Blackwell Publishing.)

7. Colony PCR identification of recombinant DNA using a pair of plasmid DNA specific primers using PCR SuperMix (Invitrogen).
8. To make further confirmation of recombinant plasmid DNA, transformants are picked and grown overnight for preparation of plasmid DNA by using QIAprep Miniprep Kit (Qiagen).
9. Recombinant DNAs are analyzed and confirmed by DNA sequencing using a pair of vector-specific primers. The resulting recombinant plasmids constructed in this example are designated as pYJ20013, pYJ20014, and pYJ20015, respectively (**Fig. 2**).

3.2. Phenotype Screening of fabI *Antisense Strains*

3.2.1. Preparation of S. aureus *Electrocompetent Cells*

S. aureus electrocompetent cells are first prepared as previous described ***(10)***. Briefly, a total of 10 mL overnight culture of RN4220/tetA, an *S. aureus* laboratory strain, is inoculated in 500 mL TSB medium and incubated at 37°C with shaking until $OD_{600\,nm}$ 0.4 to 0.5. The bacterial cells are then harvested by centrifugation at 8000 rpm for 10 min at 4°C and rinsed four times in ice-cold 0.5 M sucrose with 0.5, 0.25, 0.125, and 0.0625 times of the original bacterial culture. The cells are then resuspended with 2 mL of 10% glycerol, aliquoted, and stored in a –80°C freezer.

3.2.2. Electroporation of Antisense Constructs into S. aureus

These antisense constructs (such as pYJ20013, pYJ20014, and pYJ20015) and an empty vector (pYH4) are subsequently electroporated into 50 µL of *S. aureus* RN4220/tetA competent cells at 1.8 kV, 100 Ω resistance, and 25 µF capacitance using the Bio-Rad Gene Pulser unit. Electro-transformants are spread for single colonies on TSA-Erm (5 µg/mL) plates.

3.2.3. Screening Colonies During Induction of Antisense RNA Expression on TSA Solid Plates

In order to screen for colonies with conditional lethal phenotype or growth defects, transformants are duplicated onto TSA-Erm plates in the presence or absence of inducer Tc (1 µg/mL) and incubated overnight at 37°C. As shown in **Figure 2**, all *S. aureus* stains grew normally in the absence of Tc. In contrast, *fabI* antisense strains JSB20013 and JSB20014 did not grow at all in the presence of Tc, suggesting that they have lethal phenotype during the induction of either of these two *fabI* antisense fragments. Therefore, the *fabI* is an essential gene in *S. aureus* because inhibition of *fabI* product by JSB20013 and JSB20014 antisense fragments resulted in complete inhibition of growth. In addition, colonies containing pYJ20015 were much smaller in the presence of Tc than those in the absence of it. Therefore, pYJ20015 confers a defective phenotype upon Tc induction. Note that different antisense fragments within a gene cause different phenotypes during induction of antisense RNA. Control strains carrying parental plasmid (pYH4) grew normally with or without Tc induction.

3.3. *Quantitative Titration of Expression of Essential Genes* In Vitro

To study essential genes in a quantitative manner, growth inhibition of strains JSB20013, JSB20014, JSB20015, and the control strain RN4220/tetA/pYH4 are examined in a liquid medium containing various concentrations of Tc (**Note 4**).

1. To do so, incubate the *S. aureus* antisense strains at 37°C overnight in TSB containing 5 µg/mL Erm.
2. Dilute overnight cultures to ~10^4 colony-forming units (CFU)/mL with Erm-TSB and Tc at concentrations of 0, 10, 50, 100, 250, 500, 750, or 1000 ng/mL.
3. The cell growth was monitored for 18 h at 37°C by measuring the optical densities of suspensions at 600 nm every 15 min with 1 min mixing before each reading using a microtiter plate reader SpectraMax plus384 (Molecular Devices).

The control strain, RN4220/tetA/pYH4, did not show any significant difference with various concentrations of Tc (**Fig. 3D**). However, *fabI* antisense strains JSB20013, JSB20014, and JSB20015 (**Fig. 3A–C**) grew significantly more slowly in the presence of Tc (≥250 ng/mL). Consistent with previous results of phenotypic screenings on TSA plates, JSB20013 and JSB20014 had similar growth characteristics and their growth was completely inhibited when the concentration of Tc was 500 ng/mL or higher (**Fig. 2** and **Fig. 3B**). However, the JSB20015 strain was still able to grow slowly at 1000 ng/mL of Tc (**Fig. 3C**). Therefore, these findings have demonstrated that not all antisense RNAs are equally effective in preventing gene expression. It is necessary to create several antisense expression constructs for testing essentiality of each gene.

3.4. *Quantitative Titration of Expression of Essential Genes* In Vivo

This regulated antisense system also provides a unique tool for determining gene essentiality during infection as the tetracycline is available in many body compartments after oral dosing. We have chosen a murine model of hematogenous pyelonephritis as it results in a localized kidney infection from which bacteria are readily recovered (*13, 14*).

3.4.1. *Construction of* yhdO *and* ybcD *Antisense Strains*

To conduct quantitative titration of essential genes *in vivo*, antisense constructs YJ2-8 and YJ3-5 carrying essential *S. aureus yhdO* and *ybcD* genes, respectively, and an empty vector YJ335 were first electroporated into *E. coli* DH10B as described previously and were then electroporated into an *S. aureus* RN4220/tetA. Finally, these constructs were electroporated into a *S. aureus* clinical strain WCUH29 ready for animal infection studies.

3.4.2. Preparation of S. aureus Culture for Inoculation

S. aureus strains YJ335, YJ2-8, and YJ3-5 were harvested from 1 mL of stationary-phase culture, washed once with 1 mL of phosphate-buffered saline (PBS), and diluted to an absorbance at 600 nm of 0.2. These bacterial suspensions were diluted and plated onto TSA-Erm plates for determination of viable CFU.

3.4.3. Infection of Mouse via Tail Vein Inoculation

Five CD-1 female mice (25 g body weight) per group were infected with ~10^7 CFU of bacteria via an intravenous infection of 0.2 mL of bacterial suspension into the tail vein using a tuberculin syringe.

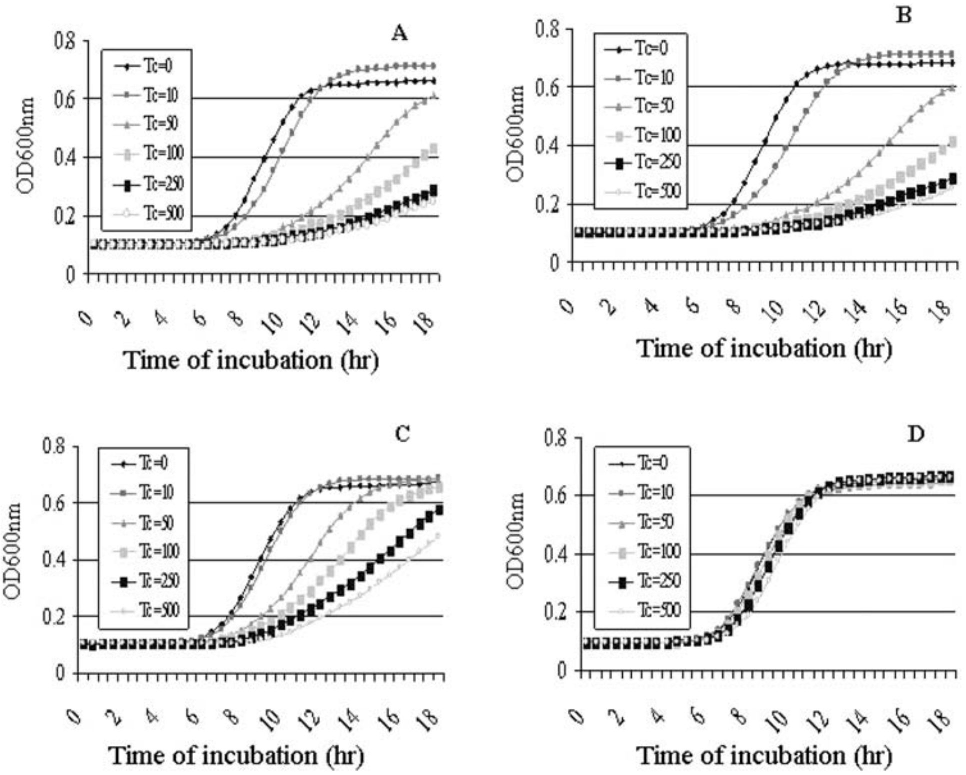


Fig. 3. Growth inhibition in the *fabI* antisense strains at different levels of antisense RNA transcription. Growth curves of **(A)** JSB20013, **(B)** JSB20014, **(C)** JSB20015, and **(D)** control strain RN4220/*tetA* carrying pYH4 were monitored for 18 h at 37°C by using a microtiter plate reader in TSB containing 5 μg/mL of Erm and varying concentrations of inducer Tc (0 to 500 ng/mL). (Reprinted from Ref. 12 by permission of Blackwell Publishing.)

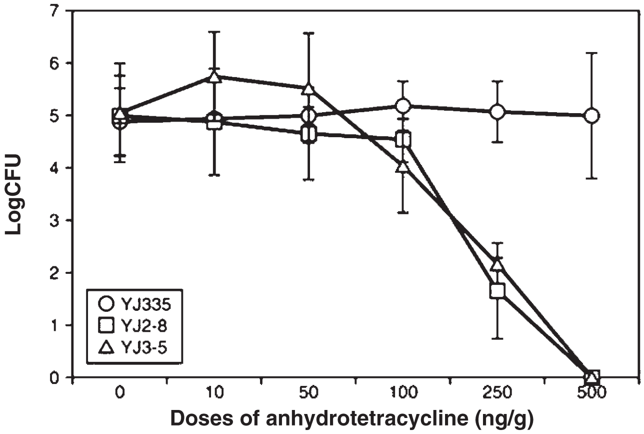


Fig. 4. Quantitative titration of expression of essential genes *in vivo*. CD-1 female mice were infected with ~10^7 CFU of *S. aureus* strain (YJ335, YJ2-8, and YJ3-5) through an intravenous injection. Various doses of ATc were given orally to infected mice on days 1, 2, and 3 after infection. The mice were killed 2h after the last dose of ATc induction, and kidneys were aseptically removed and homogenized in 1mL PBS for enumeration of viable bacteria. Reprinted from Ref. ***14*** by permission of AAAS.

3.4.4. Induction of Antisense RNA Expression During Infection

Different doses of inducer, anhydrotetracycline (ATc; a nonantibiotic analogue of tetracycline), were given orally in 0.2mL of doses containing Erm (5μg/g body weight) on days 1, 2, and 3 after infection.

3.4.5. Recovery of Bacteria from Infected Kidneys

The mice were sacrificed by carbon dioxide overdose 2h after the last dose of ATc induction. Kidneys were aseptically removed and homogenized in 1mL PBS for enumeration of viable bacteria by plating diluted bacteria on TSA plates in the presence of Erm (1μg/mL). As a control, ~5 × 10^5 CFU of YJ335 was recovered from infected kidneys at day 3 either in the presence or absence of ATc induction (**Fig. 4**). Similar numbers of antisense strains (YJ2-8 and YJ3-5) were recovered from infected kidneys without induction. However, no antisense mutants were collected after induction of antisense using 0.5μg/mL ATc/g body weight. Therefore, essential genes can be studied in the context of a titratable, conditional phenotype in a relevant model of infection.

In conclusion, the regulated antisense system described here offers a comprehensive genomic approach to identify and characterize essential genes in *S. aureus* both *in vitro* and *in vivo*.

Notes

1. PCR primer design should avoid RNA secondary structures and include antisense fragments near the ATG start codon. The size of antisense RNA should be in the 300- to 800-bp range. It is better to use computer software such as Lasergene Primerselect (DNASTAR) for PCR primer design.

2. With rapid advancements in the PCR machine design and its reagents, multiple PCR reagents and PCR machines can be utilized to generate PCR fragments.
3. To construct an antisense library, it is more efficient to pool PCR products before restriction enzyme digestions. The digested PCR products are purified by using PCR purification kit (Qiagen). The digested and pooled DNA fragments can be ligated to a digested vector in the 3:1 ratio. Plasmids can be prepared by using a QIAprep 96 Turbo Miniprep Kit (Qiagen) according to the manufacturer's instructions. Unique constructs then can be obtained by analysis of DNA sequencing results of a larger number of plasmids.
4. Quantitative titration of essential genes *in vitro* can be conducted in Erm-TSA plates containing various concentrations of Tc instead of a liquid culture if a microtiter plate reader is not available.

References

1. Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811.
2. Kernodle, D. S., Voladri, R. K. R., Menzies, B. E., Hager, C. C., and Edwards K. M. (1997) Expression of an antisense *hla* fragment in *Staphylococcus aureus* reduces alpha-toxin production *in vitro* and attenuates lethal activity in murine model. *Infect. Immun.* **65**, 179–184.
3. Wagner, E. G., and Simons, R. W. (1994) Antisense RNA control in bacteria, phages, and plasmids. *Annu. Rev. Microbiol.* **48**, 713–742.
4. Good, L., and Nielsen, P. E. (1998) Antisense inhibition of gene expression in bacteria by PNA targeted to mRNA. *Nat. Biotechnol.* **16**, 355–358.
5. Stieger, M., Wohlgensinger, B., Kamber, M., Lutz, R., and Keck, W. (1999) Integrational plasmids for the tetracycline-regulated expression of genes in *Streptococcus pneumoniae*. *Gene* **226**, 243–251.
6. Geissendorfer, M., and Hillen, W. (1990) Regulated expression of heterologous genes in *Bacillus subtilis* using the Tn10 encoded tet regulatory elements. *Appl. Microbiol. Biotechnol.* **33**, 657–663.
7. Ji, Y. (2002) The role of genomics in the discovery of novel targets for antibiotic therapy. *Pharmacogenomics* **3**, 315–323.
8. Yin, D., and Ji, Y. (2002) Genomic analysis using conditional phenotypes generated by antisense RNA. *Curr. Opin. Microbiol.* **5**, 330–333.
9. Yin, D., Fox, B., Lonetto, M. L., Etherton, M. R., Payne, D. J, Holmes, D. J., et al. (2004). Identification of antimicrobial targets using a comprehensive genomic approach. *Pharmacogenomics* **5**, 101–113.
10. Ji, Y., Woodnutt, G., Rosenberg, M., and Burnham, M. K. R. (2002) Identification of essential genes in *Staphylococcus aureus* using inducible antisense RNA. *Methods Enzymol.* **358**, 123–128.
11. Forsyth, R. A., Haselbeck, R. J., Ohlsen, K. L., Yamamoto, R. T., Xu, H., Trawick, J. D., et al. (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol. Microbiol.* **43**, 1387–1400.
12. Ji, Y., Yin, D., Fox, B., Holme, D. J., Payne, D., and Rosenberg, M. (2004) Validation of antibacterial mechanism of action using regulated antisense RNA expression in *Staphylococcus aureus*. *FEMS Microbiol Lett.* **231**, 177–184.

13. Ji, Y., Marra, A. Rosenberg, M., and Woodnutt, G. (1999) Regulated antisense RNA eliminates alpha-toxin virulence in *Staphylococcus aureus* infection. *J. Bacteriol.* **181**, 6585–6590.
14. Ji, Y., Zhang, B., Van Horn, S. F., Warren, P., Woodnutt, G., Burnham, M. K. R., and Rosenberg, M. (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science* **293**, 2266–2269.

20

Techniques for the Isolation and Use of Conditionally Expressed Antisense RNA to Achieve Essential Gene Knockdowns in *Staphylococcus aureus*

Allyn Forsyth and Liangsu Wang

Summary

This chapter provides methods and insights into the use of antisense RNA as a molecular genetic tool. Posttranscriptional inhibition of specific gene expression can be achieved by antisense RNA fragments under control of a conditional promoter. Effective titration of gene expression can cause an apparent null mutation or can be modulated to levels of interest in comparison with wild type. Validation of antisense RNA can be achieved by both RNA and protein quantitation techniques. Applications include phenotypic studies of genes in response to specific stimuli, environments, or the contribution of genes in regulatory networks. This chapter will focus on shotgun-cloned antisense for comprehensive gene identification and cell-based hypersensitivity assays for antibiotic screening. Antisense RNA strategies have high utility when the target gene is essential for survival and needs to be compared with wild type.

Key Words: antibiotics; antisense RNA; cell-based assays; drug screening; essential genes; growth inhibition; *Staphylococcus aureus*.

1. Introduction

Naturally occurring antisense RNA control of gene expression and global gene regulation has been demonstrated in a broad range of organisms (*1*). In *Escherichia coli*, these regulatory RNAs can repress or activate translation and protect or degrade mRNAs via base pairing with the target transcripts (*2*). In fact, more than 60 noncoding small RNA genes have been identified in *E. coli* (*3, 4*). More recently, there has been an increased use of artificial antisense RNA for bacterial gene discovery and the study of gene expression (*5–9*).

The mechanisms of natural and artificially expressed antisense RNAs are varied and are commonly reported to be mRNA destabilization (*1, 4, 10*). Key features of an effective antisense RNA are its persistence and accessibility to participate in RNA-RNA duplexes (*11*), which are features that are not readily predictable. While attempting to

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

saturate a given gene with antisense RNA, it has been observed that some regions will be repeatedly recovered as effective inhibitors, whereas alternate zones will not generate inhibitory antisense constructs (*[5, 12]*; *see* **Chapter 19**). The differing efficacy of various antisense RNAs for a given gene can obscure the functional difference between a marginally effective antisense targeting an essential gene and, for example, a highly efficacious antisense targeting a gene required merely for rapid growth. This illustrates a major advantage of the genome-wide antisense fragmentation approach. It selects for growth inhibitory antisense RNAs from large random populations without any *a priori* knowledge of what will make a good antisense RNA.

Research requiring modulation of gene activity can be approached by two main mechanisms: the use of a titratable promoter to control gene expression *(13–15)* or antisense knockdowns. Although antisense can be delivered by plasmid-born multicopy constructs or single-copy insertions, inherently it can only attenuate expression from endogenous levels. Promoter replacements, on the other hand, can be used to modulate expression either above or below endogenous levels. This allows the potential advantage to create a single strain that will exhibit either auxotrophy or sensitization in the attenuated state versus a gain of function or resistance in the overexpressing state.

Whereas both methods for expression modulation are useful and complementary, promoter replacements result in the alteration of a gene's normal expression context. This is particularly serious for genes that are differentially regulated during the cell cycle or in various growth conditions. The result is that some strains with promoter replacements are difficult to recover, are unstable, or grow substantially different than wild type. Alternatively, antisense RNA provides a genetically "wild type" state in the absence of induction with the ability to downregulate gene activity using a titratable promoter, making the approach ideal for the study of essential genes that will not tolerate traditional knock-out strategies.

Antisense inhibition is not without pitfalls. For instance, antisense destabilization of a transcript can create polar effects if the transcript is polycistronic. This creates a possibility for an antisense RNA complementary to a nonessential gene to destabilize an essential gene in the same polycistronic message. Well-designed promoter replacements specifically alter only the expression of their target transcript.

In this chapter, the methods required to identify a substantially comprehensive set of antisense RNA molecules targeting essential genes will be illustrated. First, a suitably inducible and titratable promoter and vector combination must be identified. Random fragments from the target genome are generated and are shotgun-cloned into the vector. Then the library is replica-plated onto noninducing media and screened for clones that have a no-growth phenotype on promoter-inducing media. The sensitivity of clones where conditional expression of the inserts blocks growth is verified, and sequence analysis of the inserts is used to identify approximately 70% of all expression-sensitive clones, which are putatively expressing antisense RNA *(5)*. Experimental strategies for verifying the efficacy and specificity of antisense RNA-producing clones will be outlined. Finally, antisense RNA sensitization assays will be described, where the specificity of known and unknown inhibitors of cell function can be evaluated.

2. Materials

1. *Staphylococcus aureus* strains RN450 and RN4220 (*16*).
2. *Escherichia coli* strains DH5α from Gibco-BRL (Carlsbad CA) and XL1Blue from Stratagene (La Jolla, CA).
3. An antisense expression plasmid appropriate for your organism (e.g., pEPSA5 for *S. aureus*) (**Notes 1** and **2**).
4. Qiagen HotStar HiFidelity polymerase kit (Qiagen, Valencia, CA).
5. Primers flanking the insert site of the cloning vector:
   (a) 25 µM LexL: TTCGCCAGACTATTTTGT
   (b) 25 µM XylT5: CAGCAGTCTGAGTTATAAAATAG
6. *Sma* I restriction endonuclease from New England Biolabs (Ipswich, MA).
7. Calf intestinal alkaline phosphatase from New England Biolabs.
8. Puregene Gram-positive DNA isolation kit from GENTRA Systems (Minneapolis, MN).
9. Lysostaphin (Sigma, St. Louis, MO) made at a 10× stock, 500 µg/mL in TE.
10. DNase I for digestion of genomic DNA into suitable fragments (Sigma).
11. T4-DNA polymerase for blunt-ending genomic fragments from New England Biolabs.
12. Qiaquick Gel Extraction Kit (Qiagen).
13. Qiagene PCR Turbo cleanup kit (Qiagen).
14. Plasmid Maxi Kit (Qiagen).
15. Bio-Rad GenePulser.
16. LB (*17*) as liquid media and agar plates supplemented with 0.2% glucose (LBG) where noted.
17. M9 salts (*17*).
18. Antibiotics: 100 µg/mL carbenicillin and 15 to 34 µg/mL chloramphenicol as indicated (Sigma).
19. SBS format single-well Omni plates (Nunc, Rochester, NY).
20. Omni plates with 75 mL LBG + 2% xylose agar.
21. Omni plates with 75 mL LBG agar.
22. Colony picking robot (GeneMachine, Ann Arbor, MI), or manual mechanism of arraying colonies into 384-well plates.
23. Replica plating robot (GenomicSolutions Flexys, Ann Arbor, MI) or manual grid tool for replica plating from 384- or 96-well plates to agar plates.
24. SBS standard 384-well plates (Matrix Tech Corp, Hudson, NH).
25. SBS standard 96-well plates (Matrix Tech Corp).
26. 1% ultrapure agarose in 1× TAE (Invitrogen, Carlsbad, CA).

3. Methods

The methods below are ordered into the major steps of (1) vector choice and preparation, (2) genomic isolation and preparation, (3) library construction, (4) library screening, (5) identification and validation of putative antisense RNA producing clones, (6) validation of antisense RNA mode of action and specificity, and (7) assays of utility using antisense RNA.

3.1. Vector Choice and Preparation, the Use of pEPSA5 Containing the pT5X Xylose-Inducible Promoter

An inducible and titratable promoter and vector combination must be identified for your organism (**Note 1**). The pEPSA5 *S. aureus/E. coli* shuttle vector (**Fig. 1**) is an

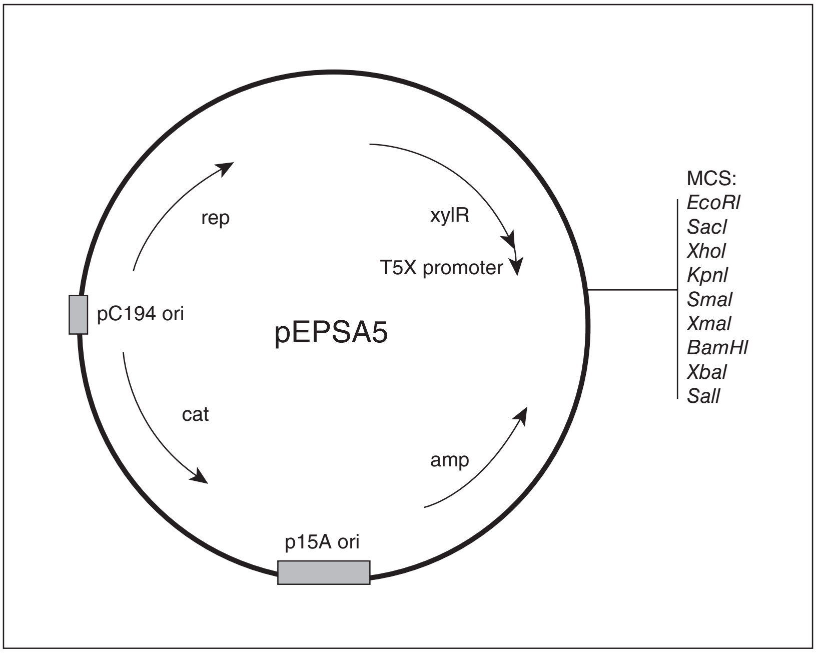


Fig. 1. Features of the pEPSA5 *S. aureus/E. coli* shuttle vector include the pC194 origin for replication and *cat* gene for selection in *S. aureus* and the low-copy-number p15A origin for replication and the *amp* gene for selection in *E. coli.* Chromosomal fragments are cloned into a polycloning site downstream of the synthetic T5X promoter, which is repressed by the *xylR* gene product and de-repressed by the addition of xylose to the growth medium.

effective vector for expression of shotgun-cloned antisense RNA in *S. aureus*. Sequences for cloning and replication in *E. coli* include the ampicillin-resistance gene of the plasmid pLEX5BA ***(18)*** and the low-copy p15a origin ***(19)*** (**Note 3**). The plasmid can be conveniently shuttled from *E. coli* into *S. aureus* by virtue of the pC194-derived replicon from pRN5548 and the chloramphenicol-resistance gene ***(20)***.

Blunt-ended fragments are cloned into the *Sma* I site of the multiple cloning region (**Note 2**). Downstream are the *rrnB* T1T2 terminators, which ensure that a discreetly sized RNA is transcribed. High-level expression of cloned inserts is achieved with the Gram-positive optimized bacteriophage T5 P_{N25} promoter ***(21)*** that has been fused downstream of the operator sequence for the *Staphylococcus xylosis* XylR repressor protein ***(22)***. This allows the use of xylose as a titrator of the pT5X promoter, and thus the vector can generate sufficient expression of antisense RNA to overwhelm endogenous levels of most target mRNAs and yet exhibits little promoter leakiness in the absence of induction allowing the maintenance of a wild-type state.

3.1.1. Preparation of Blunt-Ended pEPSA5 Vector

1. Digest pEPSA5 plasmid DNA with *Sma* I to completion.
2. Dephosphorylate the digested vector DNA with calf intestinal alkaline phosphatase (CIP).

3. The enzymes are heat-inactivated upon completion of the above reaction and the plasmid purified and quantitated for subsequent cloning steps.

3.2. Genomic DNA Isolation and Preparation

The cloning strategy employed should comprehensively represent the genome of interest in small random overlapping fragments. Restriction digests are technically simple libraries to produce but have poor randomization of fragments. Alternatively, using random mechanisms of shearing genomic DNA to be screened is more technically demanding but allows for thorough coverage when one balances the size of the fragments with the number of clones in the library. Whereas larger fragments reduce the number of clones needed to saturate a given genome, the likelihood of fragments over 800 bp containing translational sequences, which may produce dominant lethal phenotypes in the screen, goes up dramatically. Alternatively, if the library fragments become very small (<200 bp), one has to screen an excessive number of clones. A reasonable balance is a target range of 200 to 800 bp for the fragment sizes.

3.2.1. Genomic DNA Isolation

1. Grow the genomic donor strain (e.g., RN450) to mid log phase, pellet, and resuspend in the Puregene Cell Suspension Solution (Gentra Systems).
2. Add the lytic enzyme solution enhanced with 50 μg/mL lysostaphin and incubate at 37°C for 30 min.
3. All other steps are according to the manufacturer's instructions.

3.2.2. Genomic DNA Fragment Preparation

1. Genomic DNA is fragmented by standard endonuclease restriction or DNase I digestion (*17*).
2. If necessary, blunt-end the fragments with T4-DNA polymerase as described (*17*).
3. Digested DNA should be run on a 1% agarose TAE gel.
4. Fragments in the desired size range, 200 to 800 bp, are excised from the gel.
5. Purify the agarose embedded fragments using the QIAquick Gel Extraction Kit (Qiagen) according to the manufacturer's directions.

3.3. Library Construction

1. Combine fragmented and blunted genomic fragments from **Section 3.2.2** with pEPSA5 dephosphorylated blunt-end vector from **Section 3.1.1** at varying vector-to-insert ratios, ranging from 1 : 1 to 1 : 3 in a ligation mixture with T4 ligase.
2. Electroporate ligations into *E. coli* strain DH5α using a Bio-Rad GenePulser and plate on LB + 100 μg/mL carbenicillin agar plates to generate a clonal library.
3. After 16 h incubation, if greater than 1×10^6 individual colonies are generated from a transformation, they should be pooled as a library. However, take care to maintain each insert-to-vector transformation independently.
4. Purify plasmids from the library pools using Plasmid Maxi Kit (Qiagen).
5. Electroporate a portion of each library pool independently into RN4220 (*23*).
6. Plate on LBG + 15 μg/mL chloramphenicol agar plates and incubate overnight.
7. Determine the number of transformants per volume for each library (**Note 4**). Use this information to decide how much of the transformation mix to spread per plate and the number of plates you will need to sample your library.

3.4. Library Screening Using Replica-Plating Strategies

The goal of library screening is to identify those clones that produce an antisense RNA to an essential gene. This is done by observing which clones will not grow in the presence of promoter induction. In practice, approximately 1% of shotgun clones in *S. aureus* will be sensitive to induction, of which about 70% will be antisense clones to essential genes. The steps presented here will assist in the identification of desired antisense clones with suggestions for reducing the percentage of clones that are sensitive to induction but do not code for an antisense RNA (**Note 5**).

1. Transfer individual *S. aureus* transformants into 384-well plates containing 50 µL of LB + 15 µg/mL chloramphenicol medium per well using a colony picking robot (GeneMachine Gel-2-Well; **Note 6**).
2. After overnight growth at 37°C, 384-well culture plates are replica-plated with a Genomic Solutions Flexys robot onto inducing (LBG + 2% xylose) and noninducing (LBG) agar medium. Be sure to mark each pair of replica plates for later examination (**Note 7**).
3. Replica plates are incubated for approximately 10 h at 37°C (**Note 8**).
4. Examine clones on each pair of replica plates. Identify, mark, and re-array clones that do not form colonies in the presence of xylose (**Fig. 2**). Typically, plucking a colony with a sterile toothpick from the noninducing plate and inoculating 100 µL of LB with 15 µg/mL chloramphenicol in a well of a 96-well microtiter plate is convenient (**Note 9**).
5. Incubate the putative hits overnight for additional testing as described below (**Note 10**).

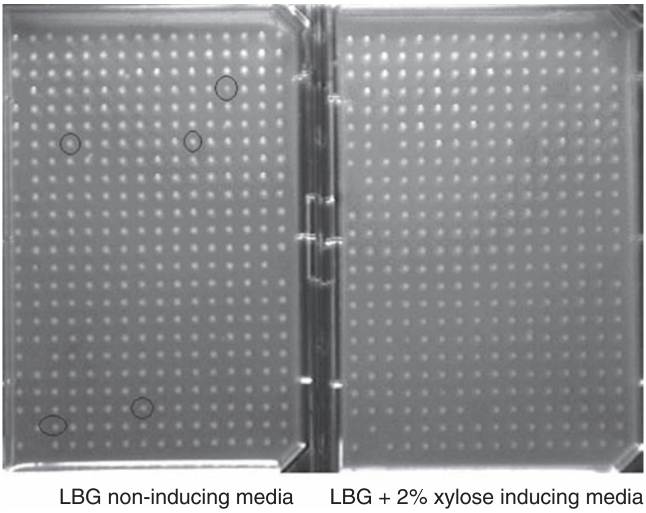


Fig. 2. Replica plating of a shotgun-cloned antisense RNA library reveals that five clones (circled) from noninducing media fail to grow on promoter-inducing media. Cultures were grown exponentially in 384-well plates and then replica plated onto agar using a 384-pin tool. Colonies that failed to grow in the presence of promoter induction were further tested to determine if they were producing antisense RNA to an essential gene.

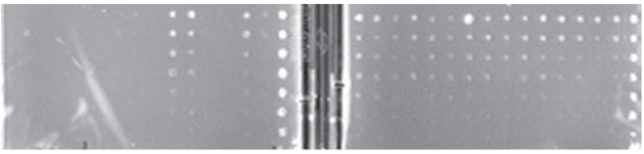


Fig. 3. Verification of clone sensitivity using replica plating of culture serial dilutions. Sixteen cultures (columns) were serially diluted 10-fold eight times and the dilutions were then replica plated (from top to bottom) onto inducing or noninducing media. Twelve clones are completely growth inhibited in the presence of xylose induction. A negative control, the pEPSA5 vector without an insert, is shown in the far left column of each plate.

3.5. Identification and Characterization of Putative Antisense RNA-Producing Clones

3.5.1. Verification of Clone Sensitivity

A simple and rapid strategy to verify sensitivity is to test serial dilutions of a culture for growth on inducing or noninducing media. This not only verifies clone sensitivity but also gives a qualitative indication of how sensitive a clone is to a specific level of inducer. For simplicity, processing of a single sample is described here (**Note 11**).

1. Collect an overnight-grown Hit plate (from **Section 3.4, step 5**) and vortex the cultures for 1 min.
2. Dilute the samples by transferring 1 μL culture into 99 μL fresh LB with 15 μg/mL chloramphenicol media and grow for 4 h at 37°C.
3. Transfer 10 μL of each culture and make 10-fold serial dilutions 8 times in 1× M9 salts in 384-well microtiter plates.
4. Replica-plate the serially diluted cultures onto inducing (LBG + 2% xylose) and noninducing (LBG) agar plates using replica plating robot or manual grid tool (**Note 12**). The grid tool is sterilized between cultures by dipping into 70% ethanol. The ethanol is then evaporated and dipped into the serial dilution plate and then onto the inducing agar plate. The tool is reinoculated from the same dilution plate and gridded onto the noninducing agar plate.
5. The plates are incubated overnight and then compared to verify clone sensitivity (**Fig. 3**).

3.5.2. Sequence Identification of Verified Sensitive Clones

Polymerase chain reaction (PCR) amplification of cloned inserts is a rapid way to generate templates for sequence analysis, which is needed to determine the insert orientation of clones.

1. From the Hit plate (made in **Section 3.4, step 5**; *see* **Note 10**), transfer 20 μL *S. aureus* overnight culture for each verified sensitive clone to a new 96-well plate. Centrifuge this 96-well plate in a tabletop centrifuge at 1800 × *g* for 10 min to pellet all the cells. Remove the media from the plate.
2. Resuspend the pelleted cells in 5 μL of 50 μg/mL lysostaphin.
3. Incubate for 30 min at 37°C.

4. Incubate at 95°C for 5 min then hold at 4°C until ready to use.
5. Add 45 µL sterile nuclease-free water to lyse the cells.
6. Transfer 2.0 µL of the lysate to a PCR premix as outlined below and amplify (**Note 13**).
   12.5 µL 2× HotStar polymerase Mix (Qiagen)
   1.0 µL LexL 25 µM
   1.0 µL XylPrimer 25 µM
   8.5 µL water
   2.0 µL of water/cell lysate mixture
   50.0 µL total volume

   The PCR program is as follows:

   (i) 95°C 15 min
   (ii) 94°C 45 s
   (iii) 54°C 45 s
   (iv) 72°C 1 min
   (v) Go to **step (ii)** 30×
   (vi) 72°C 10 min
   (vii) 4°C hold
7. PCR products are cleaned using Qiagen plate filtration system to remove primers, enzymes, and buffers.
8. The cleaned PCR products are used for sequencing to obtain the insert sequences of the sensitive clones using ABI standard sequencing protocol.

3.5.3. Determination of Insert Orientation

Bioinformatics analyses are performed to determine the original genomic locations of the sensitive clone inserts and insert orientations:

1. Homologous BLAST (blastn) comparisons of the sensitive clone insert sequences against the published *S. aureus* genomic sequence, such as N315 ***(24)***.
2. Determine the identity of the genes covered by each sensitive clone based on the similarity and the location of the top BLAST hit in the genome.
3. Determine the orientation of a cloned insert based on the relative alignment orientation between the cloned insert sequence and the gene in the genome. Those clones with inserts in the antisense orientation relative to the promoter are marked for final validation.

3.6. Validation of Antisense RNA Mode of Action and Specificity

There are many methods one can use to verify the activity and efficacy of antisense RNA and they vary in their level of technical demands and directness, including Northern blot analysis, reverse transcriptase PCR strategies, and Western blot analyses. They are not described here due to space limitations (**Note 14**). Assays for an increase in effectiveness of target inhibitors like antibiotics ***(5)*** do not demonstrate the exact molecular mechanism of action but are valuable assays for new identifying novel inhibitors of the antisense target.

3.7. Assays Using Antisense RNA

Shotgun antisense approaches generate a wealth of information regarding genes essential for growth of the target organism, and the elegance of this approach is in the ease of converting the clones directly into an assay. Because antisense expression can be modulated to levels below wild type, inhibitors of the target protein become more potent and the generated strain is referred to as *hypersensitive*. A hypersensitivity assay is most efficiently created in four steps.

3.7.1. Generating a Clone-Specific Dose-Response Curve to Antisense RNA Induction

1. Start a culture of an antisense RNA clone of interest and grow to early logarithmic growth, an OD_{600} of approximately 0.1.
2. Create a 384-well assay plate:
   (a) Titrate the final xylose concentrations in each well from 100 mM to 0 mM xylose.
   (b) Dilute the logarithmic culture to an initial inoculum equivalent to an OD_{600} = 0.0002 in LB with 34 µg/mL chloramphenicol with a final volume of 50 µL (**Note 15**).
3. Incubate the plate for 8 to 10 h in a plate reader, collecting OD_{600} reads at regular intervals (**Fig. 4**).
4. Graph the data as a growth curve or a hill plot to extrapolate the xylose IC_{50}.

3.7.2. Choosing a Target Level of Growth Inhibition

Every clone will have a unique response to xylose (**Note 16**). Measuring inhibition is best done by choosing a point in the early logarithmic growth of the no-induction control (e.g., OD_{600} = 0.1) and determining the comparable OD_{600} at the xylose level of interest at the same time point. A level of xylose that results in 20% to 80% less growth (e.g., OD_{600} = 0.02 to 0.08) is the typical range for best results (**Note 17**). For the purpose of this example, 13.5 mM xylose was chosen, which generates

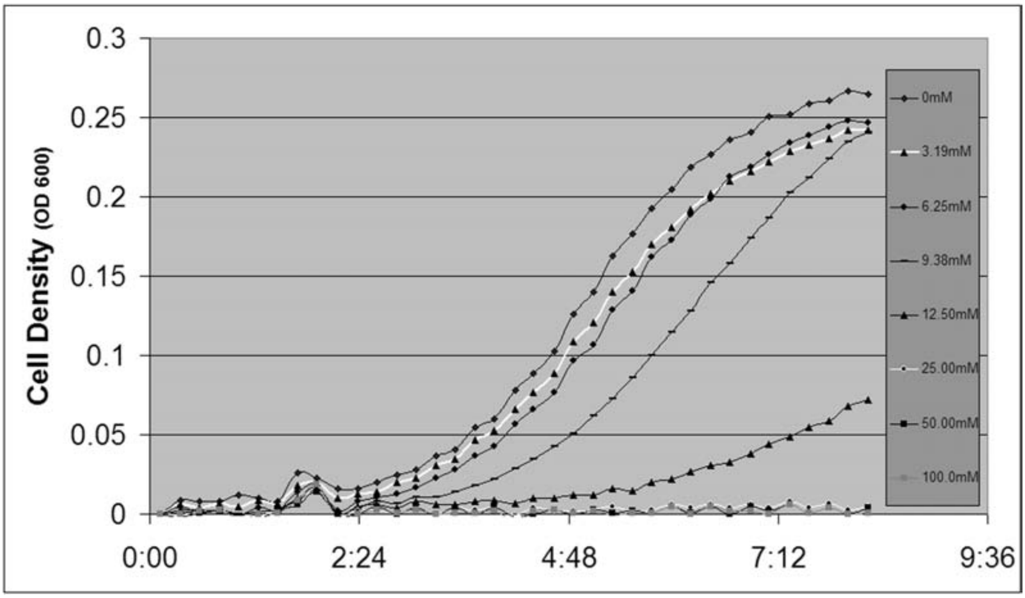


Fig. 4. A growth curve of an *rpsR* antisense clone shows decreasing growth rate with increasing levels of antisense RNA induction. An exponential culture of the *rpsR* strain was diluted with fresh LB supplemented with a range of xylose, from 0 to 100 mM, and monitored.

approximately 80% growth inhibition relative to the noninduced control for the *rpsR* antisense clone.

3.7.3. Testing the Hypersensitivity of the rpsR Antisense Strain

Once you have a chosen a level of xylose inhibition (e.g., 13.5 mM for the *rpsR* antisense strain from the previous section), generate a titration of an antibiotic known to interact with the target. The *rpsR* protein product S18 is a component of the small subunit of the ribosome. Some antibiotics are known to interact with the 30s subunit. Here spectinomycin is chosen for the test.

1. Titrate spectinomycin so that your assay plate will have a final concentration from above the MIC (approximately 1 µg/mL for *S. aureus* RN4220) down to 0, in a 7-point titration (**Fig. 5**).
2. Each well should also have an initial inoculum equivalent to an $OD_{600} = 0.0002$ of the *rpsR* strain in LB with 34 µg/mL chloramphenicol in replicate. For half the replicates, add xylose at 13.5 mM, and to the other replicates, no xylose is added.
3. Grow the cells into early log, at least an $OD_{600} = 0.1$.
4. Incubate the plate for approximately 10 h in a plate reader collecting OD_{600} reads. Take care to minimize evaporation by sealing the plate or working in a humidified environment.
5. Calculate and compare the hypersensitized IC_{50} to the noninduced IC_{50} of the strain obtained at the same OD_{600}. The *rpsR* clone induced with 13.5 mM xylose has an IC_{50} of more than fivefold lower than the noninduced control IC_{50} at the same stage of growth.

3.7.4. Testing a Panel of Antibiotics Against a Hypersensitive Strain

Prepare appropriate stocks of the antibiotics of interest. Create an antibiotic assay plate as in **Section 3.7.3** for multiple antibiotics that inhibit a variety of pathways. For each antibiotic, create replica wells without xylose as controls. To experimental wells, add xylose to a final concentration of 13.5 mM or as determined for your strain. Incubate the plate for 8 to 10 h in a plate reader collecting OD_{600} reads. Calculate the IC_{50} for each antibiotic plus and minus xylose. Calculate the fold sensitization for an antibiotic as

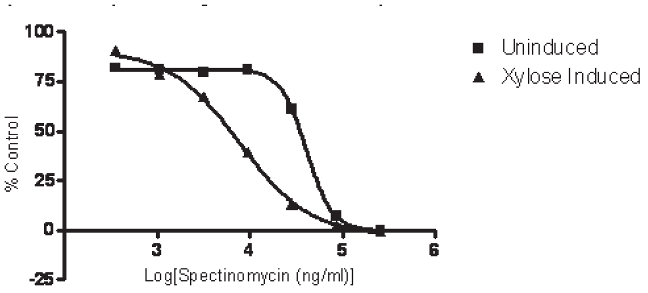


Fig. 5. A 7-point dose-response to spectinomycin tested on a *S. aureus rpsR* antisense clone without xylose and with 13.5 mM xylose to induce antisense induction. Note that spectinomycin IC_{50} of the strain shifts more than fivefold in the hypersensitized state.

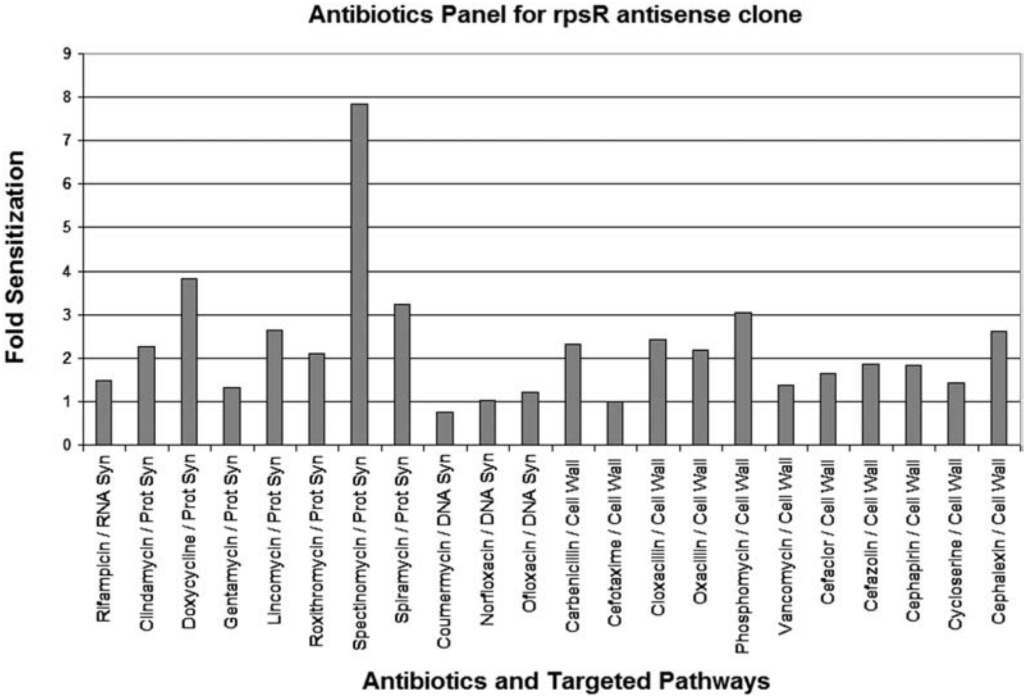


Fig. 6. Expression of the *rpsR* antisense RNA preferentially sensitizes the strain to spectinomycin (about sevenfold). In this panel, 22 antibiotics were tested against *S. aureus* with and without induction of *rpsR* antisense RNA expression. *Fold sensitization* represents the change in the IC_{50} of an antibiotic upon *rpsR* antisense RNA induction.

$$\frac{\mathrm{IC}_{50}\ \text{measured on noninduced}\ \textit{rpsR}\ \text{strain}}{\mathrm{IC}_{50}\ \text{measured on induced}\ \textit{rpsR}\ \text{strain}}$$

A full antibiotic panel using the *rpsR* strain is shown in **Figure 6**. Iterate the assays described in **Section 3.7.2** and **Section 3.7.3** until appropriate sensitization in the target-specific antibiotic is maximized without significant sensitization in the nontarget antibiotics.

4. Conclusion

The generation of antisense clones via whole-genome shotgun cloning can be a rapid mechanism of identifying essential genes. Unlike alternative screens for essential genes, antisense clones can be used in a variety of assays to elucidate gene function. In the quest for novel antibiotic discovery, antisense clones provide a graceful mechanism of converting essential gene information into practical assays and screens. For targets of known antibiotics, a hypersensitized assay allows one to identify new inhibitors of a target that is known to be druggable. Antibiotic analogues can be quickly tested in a hypersensitized assay against antibiotic panels to identify the specificity of the compound. Ideally, antisense screens are collaborated by more expensive and time-consuming biochemical assays. If an antibiotic panel assay for a gene of unknown function is generated, the sensitized antibiotics may provide clues to the pathway of that gene, and inhibitors identified in screening can allow the progression of structure-function tests and the development of a biochemical assay. The speed and flexibility of antisense assays provide a valuable tool for gene and drug discovery.

Notes

1. There are many academic choices for both Gram-negative (***25–29***) and Gram-positive bacterial promoter and vector systems (***13–14, 30–33***).
2. Common expression vectors will include translational start sequences upstream of the cloning region, which will need to be removed. The presence of any real or cryptic translational start sequences at the 5′ end of the message will allow translation of otherwise nontranslated messages and the possibility of dominant lethal peptides.
3. Repression of the T5X promoter in *E. coli* is inefficient, which can lead to selective loss of potential antisense-producing clones in *E. coli*. This is minimized by reducing the copy number in *E. coli* and allowing very few replication cycles upon transforming the shotgun library into *E. coli*.
4. The general quality of the library can also be evaluated at this stage. Ideally, the library should represent all regions of the chromosome and should have very few duplicate clones. Duplicate clones can arise during the transformation and amplification in the *E. coli* host. Additionally if insert/vector ratios were too high, multiple noncontiguous fragments of the chromosome may be cloned in tandem. If this is observed, choose a ligation with a lower insert-to-vector ratio. PCR amplification and sequencing of a representative number of inserts is a convenient way to assess the library quality.
5. Phenotypes sensitive to induction may develop due to a number reasons unrelated to antisense RNA expression, including dominant lethal peptides, nonsense peptides produced from cryptic translation sequences, or antisense RNA targeting nonessential genes that are either growth important or are in an operon with an essential gene. It is also worth noting that fragments with multiple inserts frequently are induction-sensitive, which is why generating quality libraries is so important (**Note 4**).
6. Colonies can be picked by hand depending on the investigator's timeline and the number of colonies to be screened. Additionally, direct velveteen replica plating onto the LB or LBG (inducing) plates is straightforward. Velveteen replica-plating allows for a low-cost screening mechanism but will obscure some phenotypes.
7. Handheld gridding tools are less expensive but not as reliable as robotic systems.
8. Robotic replica-plating or spotting of cultures on agar surfaces is best done with cultures that are fresh and not yet in stationary phase. Overgrown cultures will result in the deposition of a large number of cells on the inducing medium and may obscure phenotypes of interest. Experiment with initial inoculums and incubations times to optimize.
9. Inhibitory phenotypes can vary in intensity from no growth to microcolony formation in the presence of induction. It is probably better to be inclusive and note phenotypes until after a secondary verification of sensitivity is performed. The information will be useful in choosing which of several clones may suit the investigator's needs.
10. The hit plate of putative antisense clones is valuable. It is desirable to store these samples for future reference and testing. Addition of glycerol to a final concentration of 25% allows for long-term storage at −70°C.
11. To process multiple clones simultaneously, an 8-channel pipette or other robotic tool can be used to scale the suggested steps appropriately.

12. The diameter of the grid tool pins and viscosity of the fluid will determine the volumes transferred. For instance, 0.015-inch-diameter pins transfer approximately 100 nL.
13. Examining 5 μL of each PCR sample on a gel to assess the quality of the amplicons provides information on the average size of your inserts. If this examination reveals low yields (<25 ng/μL), small bands (<300 bp), or multiple bands or smears, troubleshoot the library, cell lysis (**Section 3.5.2, step 5**), or PCR steps accordingly.
14. Northern blot analysis has been used effectively to examine the mechanism of antisense action ***(34)*** and the location of mRNA cleavage ***(10)***. Verification could also be obtained by reverse transcriptase PCR strategies ***(5)***. Alternatively, a Western blot analysis can reveal if there is a decrease in the target protein with increasing antisense expression ***(34)***.
15. The initial inoculum should be prepared by diluting an early log culture with a measurable optical density. Typically, a dilution of 500-fold to 1000-fold will be required. The goal is to allow for 7 to 10 doublings of the initial inoculum in the presence of antisense induction before the culture enters stationary phase. Chloramphenicol selection for the plasmid must be maintained at all times. Cells carrying the plasmid in the induced state will be under pressure to reduce plasmid copy number. Always start cultures from noninduced cells.
16. Clones with steep xylose curves (steep-hill slopes) will change their sensitization more dramatically to small variances in xylose and tend to generate more variable data in screening situations. If multiple clones are available, test several to determine which are appropriate for your needs.
17. High levels of antisense inhibition (>50%) can cause dramatic depletions of the target protein and may significantly alter the sensitivity of the cell to a wide variety of inhibitors. Low levels of inhibition (<50%) rarely show nonspecific sensitization but may change the IC_{50} of an antibiotic by less than 2×. Experimentation with the clone of interest is recommended.

References

1. Storz, G. (2002) An expanding universe of noncoding RNAs. *Science* **296**, 1260–1263.
2. Gottesman, S. (2004) The small RNA regulators of *Escherichia coli*: roles and mechanisms. *Annu. Rev. Microbiol.* **58**, 303–328.
3. Hershberg, R., Altuvia, S., and Margalit, H. (2003) A survey of small RNA-encoding genes in *Escherichia coli. Nucleic Acids Res.* **31**, 1813–1820.
4. Kawano, M., Reynolds, A., Miranda-Rios, J., and Storz, G. (2005) Detection of 5′- and 3′-UTR-derived small RNAs and *cis*-encoded antisense RNAs in *Escherichia coli. Nucleic Acids Res.* **33**, 1040–1050.
5. Forsyth, R. A., Haselbeck, R. J., Ohlsen, K. L., Yamamoto, R. T., H. Xu, Trawick, J. D., et al. (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol. Microbiol.* **43**, 1387–1400.
6. Moreno, R., Hidalgo, A., Cava, F., Fernandez-Lafuente, R., Guisan, J. M., and Berenguer, J. (2004) Use of an antisense RNA strategy to investigate the functional significance of Mn-catalase in the extreme thermophile *Thermus thermophilus*. *J Bacteriol.* **186**, 7804–7806.
7. Bouazzaoui, K., and Lapointe, G. (2005) Use of antisense RNA to modulate glycosyltransferase gene expression and exopolysaccharide molecular mass in *Lactobacillus rhamnosus*. *J. Microbiol. Methods.* Epub. Aug 18, 2005.
8. Blokpoel, M. C., Murphy, H. N., O'Toole, R., Wiles, S., Runn, E. S., Stewart, G. R., et al. (2005) Tetracycline-inducible gene regulation in mycobacteria. *Nucleic Acids Res.* **33**, e22.

9. Wang, B., and Kuramitsu, H. K. (2005) Inducible antisense RNA expression in the characterization of gene functions in *Streptococcus mutans*. *Infect. Immun.* **73**, 3568–3576.
10. Young, K., Jayasuriya, H., Ondeyka, J. G., Herath, K., Zhang, C., Kodali, S., et al. (2006) Discovery of FabH/FabF inhibitors from natural products. *Antimicrob. Agents Chemother.* **50**, 519–526.
11. Zeiler, B. N., and Simons, R. W. (1998) In: Simons, R. W., ed. *RNA Structure and Function*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, pp. 437–464.
12. Ji, Y., Zhang, B., Van Horn, S. F., Warren, P., Woodnutt, G., Burnham, M. K., and Rosenberg, M. (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science* **293**, 2266–2269.
13. Eichenbaum, Z., Federle, M. J., Marra, D., de Vos, W. M., Kuipers, O. P., Kleerebezem, M., and Scott, J. R. (1998) Use of the lactococcal *nisA* promoter to regulate gene expression in gram-positive bacteria: comparison of induction level and promoter strength. *Appl. Environ. Microbiol.* **64**, 2763–2769.
14. Bateman, B. T., Donegan, N. P., Jarry, T. M., Palma, M., and Cheung, A. L. (2001) Evaluation of a tetracycline-inducible promoter in *Staphylococcus aureus in vitro* and *in vivo* and its application in demonstrating the role of *sigB* in microcolony formation. *Infect. Immun.* **69**, 7851–7857.
15. Kamionka, A., Bertram, R., and Hillen, W. (2005) Tetracycline-dependent conditional gene knockout in *Bacillus subtilis*. *Appl. Environ. Microbiol.* **71**, 728–733.
16. Novick, R. P. (1990) The Staphylococcus as a molecular genetic system. In: Novick, R. P., ed. *Molecular Biology of the Staphylococci*. New York: VCH Publishers, pp. 1–40.
17. Sambrook, J., Fritsch, E., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
18. Krause, M., Ruckert, B., Lurz, R., and Messer, W. (1997) Complexes at the replication origin of *Bacillus subtilis* with homologous and heterologous DnaA protein. *J. Mol. Biol.* **43**, 365–380.
19. Diederich, L., Roth, A., and Messer, W. (1994) A versatile plasmid vector system for the regulated expression of genes in *Escherichia coli*. *Biotechniques* **43**, 916–923.
20. Novick, R. P. (1991) Genetic systems in staphylococci. *Methods Enzymol.* **43**, 587–636.
21. LeGrice, S. F. J. (1990) Regulated promoter for high-level expression of heterologous genes in *Bacillus subtilis*. *Methods Enzymol.* **43**, 201–214.
22. Schnappinger, D., Geissdorfer, W., Sizemore, C., and Hillen, W. (1995) Extracellular expression of native human anti-lysozyme fragments in *Staphylococcus carnosus*. *FEMS Microbiol. Lett.* **43**, 121–127.
23. Schenk, S., and Laddaga, R. A. (1992) Improved method for electroporation of *Staphylococcus aureus*. *FEMS Microbiol. Lett.* **43**, 133–138.
24. Kuroda, M., Ohta, T., Uchiyama, I., Baba, T., Yuzawa, H., Kobayashi, I. et al. (2001) Whole genome sequencing of methicillin-resistant *Staphylococcus aureus*. *Lancet* **43**, 1225–1240.
25. de Boer, H. A., Comstock, L. J., and Vasser, M. (1983) The tac promoter: a functional hybrid derived from the *trp* and *lac* promoters. *Proc. Natl. Acad. Sci. U.S.A.* **80**, 21–25.
26. Elvin, C. M., Thompson, P. R., Argall, M. E., Hendry, P., Stamford, N. P., Lilley, P. E., and Dixon, N. E. (1990) Modified bacteriophage lambda promoter vectors for overproduction of proteins in *Escherichia coli*. *Gene* **87**, 123–126.
27. Guzman, L. M., Belin, D., Carson, M. J., and Beckwith, J. (1995) Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J. Bacteriol.* **177**, 4121–4130.

28. Yanisch-Perron, C., Vieira, J., and Messing, J. (1985) Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* **33**, 103–119.
29. Ehrt, S., Guo, X. V., Hickey, C. M., Ryou, M., Monteleone, M., Riley, L. W., and Schnappinger, D. (2005) Controlling gene expression in mycobacteria with anhydrotetracycline and Tet repressor. *Nucleic Acids Res.* **33**, e21–e21.
30. de Ruyter, P. G., Kuipers, O. P., Beerthuyzen, M. M., van Alen-Boerrigter, I., and de Vos, W. M. (1996) Functional analysis of promoters in the nisin gene cluster of *Lactococcus lactis*. *J. Bacteriol.* **178**, 3434–3439.
31. Geissendorfer, M., and Hillen, W. (1990) Regulated expression of heterologous genes in *Bacillus subtilis* using the Tn10 encoded tet regulatory elements. *Appl. Microbiol. Biotechnol.* **33**, 657–663.
32. Kim, L., Mogk, A., and Schumann, W. (1996) A xylose-inducible *Bacillus subtilis* integration vector and its application. *Gene* **181**, 71–76.
33. Kuipers, O. P., Beerthuyzen, M. M., de Ruyter, P. G., Luesink, E. J., and de Vos, W. M. (1995). Autoregulation of nisin biosynthesis in *Lactococcus lactis* by signal transduction. *J. Biol. Chem.* **270**, 27299–27304.
34. Yin, D., Fox, B., Lonetto, M. L., Etherton, M. R., Payne, D. J., Holmes, D. J., et al. (2004) Identification of antimicrobial targets using a comprehensive genomic approach. *Pharmacogenomics* **5**, 101–113.

21

Introduction of Conditional Lethal Amber Mutations in *Escherichia coli*

Christopher D. Herring

Summary

A method is described for generating conditional lethal mutations in essential genes in *Escherichia coli*. In this procedure, amber stop codons are introduced as "tagalong" mutations in the flanking DNA of a downstream antibiotic-resistance marker by lambda Red recombination. The marker is removed by expression of I-*Sce*I homing endonuclease, leaving a markerless mutation. The mutants then depend upon expression of a suppressor transfer RNA (tRNA) for survival, which is expressed under control of the arabinose promoter on a high-copy-number plasmid.

Key Words: conditional lethal; *Escherichia coli*; essential genes; lambda Red; markerless; mutagenesis; suppressor.

1. Introduction

Conditional lethal mutants are useful because they conclusively establish the essentiality of a gene and can be used to study the function of that gene. They are important in genomics-based drug discovery for target validation and prioritization and can be used to screen compound libraries and determine the mode of action.

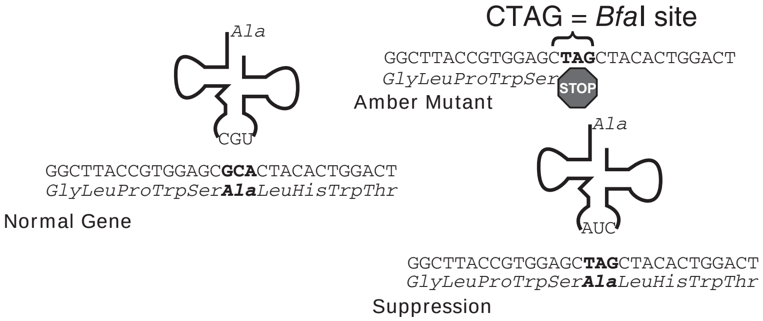
Several approaches to conditional mutagenesis can be considered. The ideal method to make conditional lethal mutations will be predictable and high-throughput and will be directed to specific genes. It should result in immediate and complete inactivation of the gene yet be reversible to wild type (WT) activity, should not cause polar effects on other genes, and should retain the native promoter so that transcriptional levels are identical to that in the wild type. Temperature-sensitive (TS) mutations offer direct and sometimes reversible control of protein function, but it is extremely difficult to design proteins with a TS phenotype. In some cases, isolation of TS alleles may not be possible (*1*) or activity may be suboptimal even at permissive temperature. Another approach is to place the target gene under the control of an experimentally controllable promoter either on a plasmid (*2*) or in the genome (*3*). This allows turning gene expression on and off over a large dynamic range, but the natural expression level and regulation of

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

the gene are overridden by the inducing promoter. Another approach uses antisense RNA in which short pieces of RNA are produced that interfere with translation (*[4, 5]*; see **Chapter 19** and **Chapter 20**). Antisense mutants typically are not targeted to specific genes and are made by using a random whole-genome approach.

The method described here uses conditional suppression of amber mutations through inducible expression of an amber suppressor transfer RNA (tRNA) *(6)*. The introduction of an amber stop codon into an essential gene leads to premature termination of protein synthesis and truncation of the encoded protein. If an essential domain is downstream and protein synthesis is unable to resume in frame, the cell will be unable to grow. This effect can be relieved by the expression of a suppressor tRNA, which is a tRNA with a sequence modification that allows it to recognize a stop codon and insert an amino acid in its place (**Fig. 1**) (reviewed in Ref. *7*). Of the three stop codons, the amber codon was chosen because it is the least common in *E. coli*, terminating only 326 out of 4290 originally annotated open reading frames (ORFs). The Ala2 suppressor was selected because of its high suppression efficiency and its specificity in introducing only the correct amino acid *(8)*. We also have found that expression of this suppressor causes minimal perturbation of global transcription in *E. coli* *(9)*.

Suppressor tRNAs have been widely used in studies of translation, phage biology, and protein engineering. They have been critical to our understanding of the structure-function, processing, and charging of tRNAs, as well as ribosome-tRNA interaction, polarity, codon context effects, and the elucidation of the genetic code *(10)*. Amber mutations were discovered in multiple labs, but their significance was not known at first. In 1958, bacteriophage mutants were isolated that showed a phenotype in some bacterial hosts but not in others, which Alan Campbell called "host-defective" mutants and Seymour Benzer called "ambivalent" mutations. The name "amber" mutations was first used by Dick Epstein and C.M. Steinberg in 1960 at CalTech. One evening while searching for strain-specific growth in phage T4 mutants, a graduate student named Harris Bernstein asked them if they wanted to go to a movie. Epstein convinced Bernstein to stay and help them pick plaques instead, and in exchange, they would name



CTAG = *Bfa*I site
Ala
CGU
GGCTTACCGTGGAGCGCACTACACTGGACT
GlyLeuProTrpSerAlaLeuHisTrpThr
Normal Gene
GGCTTACCGTGGAGCTAGCTACACTGGACT
GlyLeuProTrpSer
STOP
Amber Mutant
Ala
AUC
GGCTTACCGTGGAGCTAGCTACACTGGACT
GlyLeuProTrpSerAlaLeuHisTrpThr
Suppression

Fig. 1. Amber suppression with a suppressor tRNA. On the left, the DNA and protein sequence of a normal gene is shown, with a WT tRNA for alanine above. On the top right, the same gene is shown with an amber codon introduced in place of the alanine codon, forming the restriction site for *Bfa*I for easy identification. On the bottom right, an alanine suppressor tRNA with anticodon that matches the amber codon is shown with the resulting WT protein below.

the mutants after his mother. The mutations were then called "amber," which is the translation of Bernstein from German (*11*). It was later found that amber mutations occurred in a multitude of genes yet were suppressed by the same host factor. This led Benzer to hypothesize that the suppressor acted at the level of information transfer (*12*). By subsequent mutation studies, it was determined that amber mutations were caused by the triplet UAG, one of three "nonsense" codons that do not encode any amino acid (*13*). By showing that the amber codon caused premature protein termination, it was established as a "stop" punctuation signal (*14*). In keeping with the name, the other two stop codons were named "opal" and "ochre." The suppressor factors were identified as suppressor tRNAs that were present only in some strains (*15, 16*).

In previous work (*6*), a method called "gene gorging" utilized lambda Red recombination to introduce amber mutations into the *E. coli* genome. Unfortunately, this method proved to only be useful for the mutation of nonessential genes. A more indirect method called "tagalong mutagenesis" was developed for essential genes (*17*) and is presented here as a detailed protocol.

An overview of the "tagalong mutagenesis" strategy is shown in **Figure 2**. The first part of the procedure consists of designing polymerase chain reaction (PCR) primers and generating a linear DNA that can serve as a substrate for lambda Red recombination (**Section 3.1**). Lambda Red is induced in the target strain, and electrocompetent cells are produced (**Section 3.2**). The linear DNA is then electroporated into the cells, and conditional lethal mutants are generated by selection and screening (**Section 3.3**). The amber mutation is not selected for; instead, it is introduced as a "tagalong" in the DNA flanking a selectable antibiotic-resistance gene. In the last two steps, the selectable marker is removed, and the strain is cured of the plasmid carrying lambda Red (**Section 3.4**).

2. Materials

1. *Escherichia coli* strain MG1655 (other strains can be used as well, but they should not contain mutations in genes encoding recombination functions such as *recA*).
2. PCR reagents: a thermocycler, *Pfu* Turbo DNA polymerase (Stratagene, La Jolla, CA), the buffer supplied with *Pfu* Turbo, deoxynucleotide triphosphates (dNTPs) (2.5 mM each), molecular biology grade water, DNA of a plasmid containing the chloramphenicol-resistance gene (*cat*) (e.g., pACYC184), and genomic DNA from *E. coli*. The oligonucleotide primers OF105 (cam-2), ccgctcttcagatcctagggataacagggtaatttacgccccgccctgccact, and OF375 (cam-1), ctgttatccctaggcgcgcctaaatcctggtgtccctgttg, will be needed in addition to those designed in **Section 3.1**.
3. QIAquick PCR purification kit (Qiagen, Valencia, CA).
4. Agarose gel electrophoresis equipment to check PCR products.
5. Transformation and storage solution (TSS; available from Epicentre, Madison, WI) and sterile microcentrifuge tubes.
6. Plasmid pK-HT (*17*), which contains I-SceI and the lambda Red genes under control of the rhamnose promoter. This plasmid can be obtained from the lab of Fred Blattner (e-mail: mutants@genome.wisc.edu).
7. Plasmid pBAD/sup2 (*6*), which contains the Ala2 amber suppressor under control of the arabinose promoter. This plasmid can be obtained from the lab of Fred Blattner (e-mail: mutants@genome.wisc.edu).

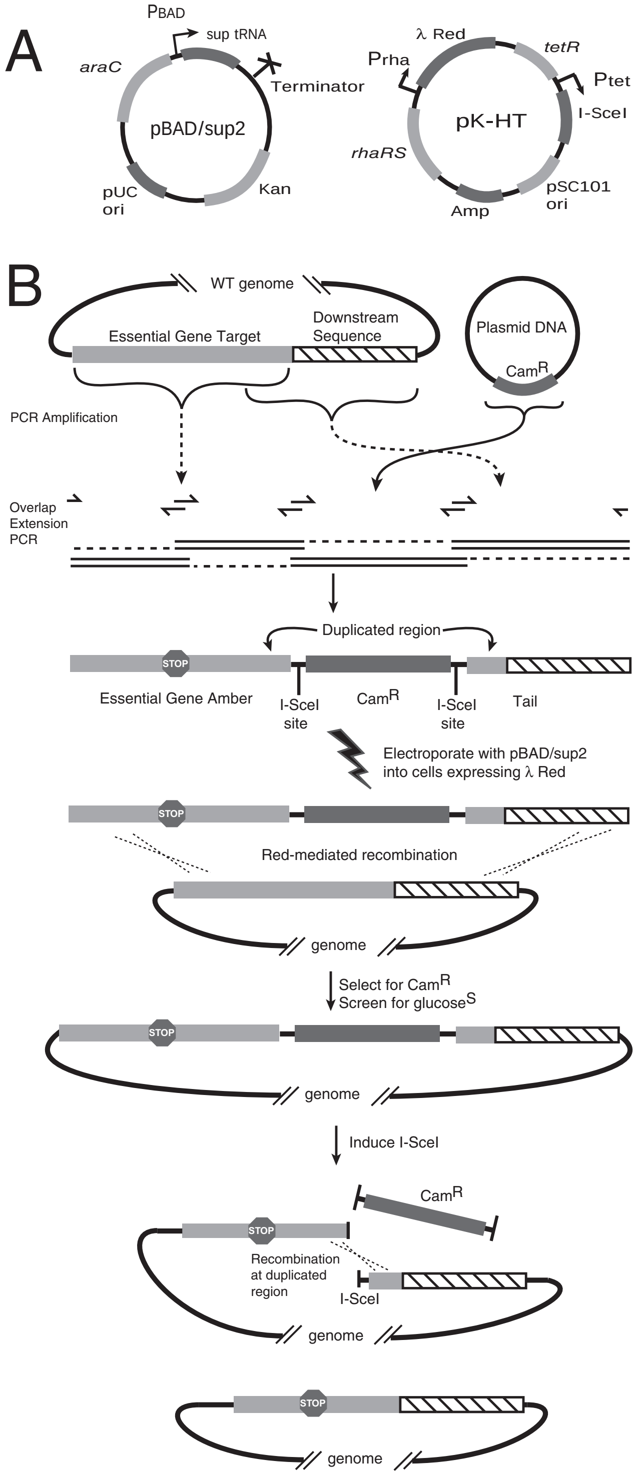


Fig. 2. **(A)** Plasmids used in tagalong mutagenesis (not to scale). **(B)** Mutagenesis strategy. A linear DNA fragment is produced from WT template genomic DNA by overlap extension PCR. The positions of primers are indicated by one-sided arrows. The PCR fusion product is electroporated into cells, and integrations resulting from a double recombination event are selected. After identification of a clone carrying the tagalong amber stop codon, the I-*Sce*I gene is induced, resulting in removal of the gene encoding Cam^{R}. Recombination within a short duplicated region leads to the generation of an amber mutant that is otherwise scarless. (Reprinted from Ref. *17* with permission from the American Society for Microbiology.)

8. Chilled, sterile deionized water and chilled, sterile 10% glycerol.
9. A refrigerated centrifuge capable of spinning 50-mL conical tubes.
10. A shaking incubator.
11. An electroporator and electroporation cuvettes.
12. Sterile toothpicks.
13. Rich defined medium (RDM), prepared as described by Neidhardt et al. (*18*) (recipe at http://www.genome.wisc.edu/functional/protocols.htm) or purchased from Teknova (Half Moon Bay, CA).
14. Liquid Luria Bertani (LB) medium.
15. SOC/Ara: Dissolve 20 g tryptone peptone, 5 g yeast extract, and 0.5 g NaCl in 1 L final volume deionized water, then add 2.5 mL 1 M KCl and adjust pH to 7.0 with NaOH. Autoclave. When the solution has cooled, add 1 mL 2 M $MgCl_2$ and 20 mL filter-sterilized 20% arabinose.
16. Antibiotics to be used at the following final concentrations: 100 ng/mL anhydrotetracycline (aTc), 100 µg/mL ampicillin (AMP), 50 µg/mL kanamycin (KAN), and 25 µg/mL chloramphenicol (CAM).
17. Sterile phosphate-buffered saline (PBS; 150 mM sodium chloride, 10 mM sodium phosphate pH 7.4).
18. 20% D-glucose (Glu), dissolved in deionized water and filter-sterilized, then used at a final concentration of 0.2%.
19. 20% L-rhamnose (Rha), dissolved in deionized water and filter-sterilized, then used at a final concentration of 0.2%.
20. 20% L-arabinose (Ara), dissolved in deionized water and filter-sterilized, then used at a final concentration of 0.2%.
21. The composition of plates is given by abbreviations of the base medium (either LB or RDM) then the antibiotics and sugars that are present. The following types of plates will be used: LB/AMP, RDM/KAN/CAM/AMP/Ara, RDM/KAN/CAM/AMP/Glu, RDM/KAN/AMP/aTc/Ara, RDM/KAN/AMP/Ara, RDM/KAN/Ara/Rha, and RDM/KAN/Ara. RDM-based plates are made by mixing (1) 500 mL of 2× concentrated RDM and sugars in sterile deionized water with (2) 13 g agar autoclaved in 500 mL deionized water. These two solutions are mixed immediately after the agar is autoclaved, then the antibiotics are added and plates are poured immediately.

3. Methods

3.1. Design of Primers/Overlap Extension PCR Amplification

Four individual pieces are amplified separately with primers designed to overlap one another (**Fig. 3**). The ORF is amplified in two halves with primers at the amino and carboxyl termini, paired with overlapping divergent primers in the middle of the gene, both containing the desired amber mutation. The gene for chloramphenicol resistance (Cam^R) is amplified with primers OF105 and OF375, each containing the recognition site for I-*Sce*I. A small part at the end of the ORF and some of the downstream sequence is amplified as a fourth “tail” fragment, with one primer ~100 nucleotides (nt) before the C-terminus (tail start) and another located ~500 nt downstream (tail end). The overlapping PCR products are then joined in a two-round overlap extension PCR. In the first round, the four primary pieces are joined to a neighboring fragment in three pairwise combinations. In the second round, the pairwise reactions are combined, and the complete fusion is amplified.

1. Primer design (**Fig. 3**). Design a primer immediately downstream of the 3′ end of the target gene in the reverse direction (C-term primer). Append the following sequence to the 5′ end: 5′-aggcgcgcctagggataacagggtaat-3′. The next primer (A-term primer) should be positioned at least 1000 bp upstream in the forward orientation. Typically, the A-term primer corresponds with the 5′ end of a gene, but for genes shorter than 1000 bp or longer than ~3000 bp, this is not practical. The amber mutation will be introduced halfway between the C-term and A-term primers. Be sure to position the A-term primer so that the amber mutation will be within the target gene and will truncate as much of the protein as possible. Next, select an alanine codon (GCX) halfway between the C-term and A-term primers to change to amber (TAG). If possible, an alanine codon preceded by a C nucleotide should be chosen so that each mutant is marked by CTAG, the relatively rare restriction site for *Bfa*I, so that mutagenesis can be easily confirmed by PCR and restriction digestion. If no alanine codon with an upstream C nucleotide is available, select one that can be changed to a C without altering the amino acid sequence. Select the nucleotides upstream of the CTAG, counting 1 for A or T and 2 for C or G, for a total of 34. This will result in a T_M of approximately 68°C for that sequence. Append CTAG to the 3′ end, then additional downstream sequence counting 1 for A or T and 2 for C or G, for a total of 16. Then, reverse complement the constructed sequence; this is primer mut-1. For the next primer, select sequence upstream of the CTAG for a total count of 16. Append CTAG to the 3′ end, then additional downstream sequence for a total count of 34. This is primer mut-2. For the "tail-start" primer, design it in the forward orientation approximately 100 bp before the 3′ end of the target gene, then append the sequence atccctaggatctgaagagcgg to the 5′ end. Finally, for the "tail-end" primer, design it 400 to 600 bp downstream of the 3′ end of the target gene in the reverse orientation.
2. Using *Pfu* Turbo DNA polymerase, PCR amplify the four primary pieces in 50-μL reactions. Piece no. 1 primers are A-term and mut-1 amplifying *E. coli* genomic DNA. Piece no. 2 primers are mut-2 and C-term amplifying *E. coli* genomic DNA. Piece no. 3 primers are OF105 and OF375 amplifying plasmid DNA containing the chloramphenicol-resistance gene. Piece no. 4 primers are tail-start and tail-end, amplifying *E. coli* genomic DNA. Purify each piece with QIAquick.
3. For fusing the four primary pieces, perform all amplifications with *Pfu* Turbo DNA polymerase. In the first round, set up three reactions combining 4 μL of each pair of neighboring pieces (pieces 1 + 2, pieces 2 + 3, and pieces 3 + 4) with *Pfu* buffer, dNTPs (2.5 mM each), and *Pfu* Turbo in a 25-μL reaction mixture and then subject it to 20 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 2 min.
4. In the second round of fusion, combine 5 μL of each unpurified product from the three pairwise combinations with 3.5 μL of *Pfu* buffer, 5 μL of deoxynucleoside triphosphates, 1 μL of *Pfu* Turbo, and 20.5 μL of H_2O and subject the reaction to five cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 6 min. Then, add a 2.5-μL volume of each of the A-term and tail-end primers (5 μM) and cycle the reactions 20 to 25 more times by the same regimen. Smears and multiple bands are often visible on agarose gels, but a strong primary product of the correct size should be observed.
5. Purify the PCR products using QIAquick (Qiagen) and measure DNA concentration using a spectrophotometer.

3.2. Preparation of Lambda Red Electrocompetent Cells

In order to introduce the desired amber mutation, it is necessary to generate cells (a) that carry the plasmid pK-HT, (b) that are competent for electrotransformation, and (c)

```
                     A-term                                                                  mut-2
             atgactgaatcttttgctcaactc                                        atccgtgggtCTAGatcgctaaacgttatccggaagg

CTGAAGATTAAACATGACTGAATCTTTTGCTCAACTCTTTGAAGAGTCC...CTGGGCCTGAAACAGCTGGGCGAAGATCCGTGGGTAGCTATCGCTAAACGTTATCCGGAAGGTACCAAACTGAC
GACTTCTAATTTGTACTGACTTAGAAAACGAGTTGAGAAACTTCTCAGG...GACCCGGACTTTGTCGACCCGCTTCTAGGCACCCATCGATAGCGATTTGCAATAGGCCTTCCATGGTTTGACTG

                                                                  cgacccgcttctaggcacccaGATCtagcgatttgc
                                                                                  mut-1



                 tail-start
ATCCCTAGGATCTGAAGAGCGgctgacgagaaagatgcaatcgc

         GCGAAAGACGAAGCTGACGAGAAAGATGCAATCGCAACTGTTAACAA...ATGGCTGAAGCTTTCAAAGCAGCTAAAGGCGAGTAATTCTCTGACTCT
         CGCTTTCTGCTTCGACTGCTCTTTCTACGTTAGCGTTGACAATTGTT...TACCGACTTCGAAAGTTTCGTCGATTTCCGCTCATTAAGAGACTGAGA

                                                                       aagtttcgtcgatttccgctcattaaTGGGACAATAGGGATCCGCGCGGA
                                                                                           C-term



...ACGTAATCCGAAGACTGGCGATAAAGTAGAACTGGAAGGAAAATACG
...TGCATTAGGCTTCTGACCGCTATTTCATCTTGACCTTCCTTTTATGC

       ctgaccgctatttcatcttgacc
              tail-end



               OF375 (cam-1)
CTGTTATCCCTAGGCGCGCCtaaatcctggtgtccctgttg

       CTCGCAGAATAAATAAATCCTGGTGTCCCTGTTGATACCGGGAAGC...CAGTACTGCGATGAGTGGCAGGGCGGGGCGTAATTTTTTTAAGGC
       GAGCGTCTTATTTATTTAGGACCACAGGGACAACTATGGCCCTTCG...GTCATGACGCTACTCACCGTCCCGCCCCGCATTAAAAAAATTCCG

                                                                     tcaccgtcccgccccgcattTAATGGGACAATAGGGATCCTAGACTTCTCGCC
                                                                                      OF105 (cam-2)
```

Fig. 3. An example of primer design for mutagenesis of the *E. coli* gene *rpsA*. Double-stranded template sequence is shown with corresponding oligonucleotide primers either above (5′ to 3′ orientation) or below (3′ to 5′ orientation). The top three lines of sequence are of the *E. coli* genome, and the bottom one is of the plasmid pACYC184. The parts of the primers that match the template are in lowercase letters, and nonmatching nucleotides are in uppercase letters.

in which the lambda Red genes have been induced, allowing recombination with linear DNA.

1. Transform *E. coli* strain MG1655 with pK-HT using TSS (Epicentre; **Note 1**). Inoculate MG1655 into 2 mL LB and incubate on a rotary shaker overnight at 37°C. Then, dilute the culture 1 : 100 into 4 mL fresh LB and grow until the optical density at 600 nm (OD_{600}) reaches 0.25 to 0.4 (approximately 2 h). Pipette 1 mL of the culture into a sterile 1-mL microcentrifuge tube and centrifuge at 10,000 × *g* for 1 min. Carefully remove the supernatant, then resuspend the cells in 100 µL of ice-cold 1X TSS. Add 2 ng of pK-HT, mix by finger flicking, then incubate on ice 10 min, at room temperature 10 min, then on ice 10 min. Pipette the cells into a 15-mL snap-cap tube containing 500 µL of LB and grow for 1 h on a rotary shaker at 37°C. Spread 20 and 200 µL on LB/AMP plates and incubate overnight at 37°C.
2. Start making electrocompetent recombinogenic cells by inoculating *E. coli* carrying pK-HT into 4 mL LB + AMP and shaking overnight at 37°C. Then, dilute the culture 1 : 500 into 200 mL LB + AMP in a 500-mL flask and grow with vigorous shaking (~200 rpm) at 37°C.
3. A sterile solution of 20% rhamnose should be added to a final concentration of 0.2% approximately 2 h before harvesting to induce the plasmid-encoded lambda Red functions.
4. When the OD_{600} reaches 0.5, pour 100 mL of cells into two sterile 50-mL screw-cap conical tubes. Centrifuge at 5000 × *g* for 10 min at 4°C. In all subsequent steps, cells should be kept as close to 0°C as possible by keeping them on ice or performing the steps in a cold room. Carefully decant supernatants, then add approximately 20 mL sterile chilled deionized water to each tube. Keeping the cells on ice, resuspend the cells by pipetting up and down, then add another 30 mL chilled water to each and mix to wash the cells. Centrifuge again, then decant, resuspend, and wash the cells in chilled deionized water as before. Centrifuge again and wash once more the same as before, but use sterile chilled 10% glycerol in place of the water. This makes a total of two washes in water and one wash in 10% glycerol. It is normal for cell pellets to become looser in later wash steps. Some loss of cells is inevitable during decanting but should be kept as small as possible. Centrifuge the cells one final time and decant the supernatant. Resuspend each pellet in 500 µL of 10% glycerol, combine the cells, then dispense 100-µL aliquots into chilled sterile 1.5-mL microcentrifuge tubes. Keeping the cells on ice, use them immediately or freeze at −80°C.

3.3. Electroporation/Growth/Screening for Lethals

1. If using frozen competent cells, defrost them by putting them on ice for approximately 30 min. Chill electroporation cuvettes and sterile 1.5-mL microcentrifuge tubes on ice. Dispense 500 µL SOC/Ara into sterile 15-mL snap-cap tubes (**Note 2**). Place 100 to 200 ng of overlap extension PCR product into a chilled microcentrifuge tube, then add 70 to 90 ng of pBAD/sup2 plasmid DNA. Mix the electrocompetent cells by finger flicking, then add 40 µL of cells to the microcentrifuge tube containing the DNA. Electroporate the DNA/cells mixture according to the instructions supplied by your electroporator manufacturer.
2. Immediately remove cells from the electroporation cuvette into a prepared tube of SOC/Ara using a sterile gel-loading pipette tip. Incubate with shaking for 1 h at 37°C. Plate the

cells on RDM/KAN/CAM/AMP/Ara plates and incubate for 16 to 24 h at 37°C (**Note 3**).

3. In order to patch a colony, pick a colony with a sterile toothpick and then make a small mark (~0.5 cm) on the surface of a series of plates, marking the plate carefully with a pen so that patches on different plates can be tracked. To identify conditional lethal mutants, patch colonies from the transformation plates on RDM/KAN/CAM/AMP/Glu and RDM/KAN/CAM/AMP/Ara plates to identify clones that can grow on Ara but not on Glu (**Note 4**).

3.4. Removal of Drug Marker with I-SceI and Plasmid Curing

The chloramphenicol-resistance gene downstream of the targeted essential gene will be removed by expressing the yeast homing endonuclease I-SceI. This enzyme cuts the two copies of its 18-bp recognition sequence introduced into the genome with primers OF105/OF375, but nowhere else in the genome. This double-strand break will be lethal to the cells unless repaired by homologous recombination, most likely occurring between the 100-bp duplications of the 3′ end of the target gene, adjacent to the I-SceI sites (**Fig. 2**). After induction of I-*Sce*I, 100- to 3000-fold fewer colonies should grow on plates relative to uninduced controls, indicating effective counterselection. In some cases, mutants displayed slightly different growth properties before and after removal of the downstream Cam^R marker, so it should be removed to avoid unintentional effects.

1. Resuspend conditional lethal mutants from the RDM/KAN/CAM/AMP/Ara patch in 1 mL sterile PBS. Spread 100 µL on a RDM/KAN/AMP/aTc/Ara plate and incubate at 30°C overnight to induce I-*Sce*I without inducing lambda Red recombinase (**Note 5**). An alternate counterselection method utilizing the tetracycline resistance gene rather than I-*Sce*I is also possible (**Note 6**).
2. Patch survivors on RDM/KAN/AMP/CAM/Ara and RDM/KAN/AMP/Ara plates and grow overnight at 37°C to confirm loss of the Cam^R marker.
3. The mutagenesis plasmid pK-HT is then removed from the conditional mutants by inducing lambda Red in the absence of selection for the plasmid (**Note 7**). Streak cells from the RDM/KAN/AMP/Ara patch onto an RDM/KAN/Ara/Rha plate and grow overnight at 37°C. Resuspend a colony in 1 mL sterile PBS, then dilute 1 : 1000 in sterile PBS. Spread 100 µL on an RDM/KAN/Ara/Rha plate and grow overnight at 37°C.
4. To identify clones that no longer carry the plasmid, patch on RDM/KAN/AMP/Ara and RDMKAN/Ara plates.
5. Confirm that the targeted essential gene contains an amber mutation by PCR amplifying from one isolate of each mutant with primers located adjacent to the ORF so that only the chromosomal locus is amplified (not the PCR product that was introduced). Purify the PCR product with QIAquick and perform Sanger sequencing.

Notes

1. Alternately, pK-HT can be transformed into electro- or chemically competent MG1655, but for transformations with supercoiled plasmid DNA, the TSS method is by far the easiest.
2. The medium SOC is often used for the recovery of cells after electrotransformation, but it contains glucose, which if used here would interfere with expression of the suppressor under

control of the arabinose promoter. Thus, a modified form of SOC is used with arabinose in place of glucose.

3. The medium seems to be important for full expression of the arabinose-inducible suppressor. For all genes, the conditional lethal clones tend to be small colonies on the transformation plates. For the gene *rpsA*, conditional lethal colonies were identifiable only after 24 h of growth. These very small colonies were too small to be detected after 16 h. The small size may be due to slow growth from incomplete suppression but may also be due to lag time before the suppressor is fully expressed. Very few conditional lethal mutants were obtained when plates were grown at 30°C. For this reason, the plasmid origin from pKD46 was corrected during construction of pK-HT (**Fig. 2**) so that it would replicate at 37°C.
4. Arabinose induces the suppressor and glucose represses it. Cells that are unable to grow on plates containing glucose require induction of the suppressor for survival. These cells are conditional lethal and contain the desired amber mutation. In the original study describing this method ***(17)***, one of the eight attempted essential genes (*ftsZ*) did not yield any conditional mutants, but for the other genes, between 2.5% and 50% of the screened colonies were conditional lethals. Polarity effects on the expression of downstream genes may be an important consideration. Amber mutations are not expected to have polar effects except in cases of *rho*-mediated termination. The difficulty mutating *ftsZ* may have been due to the presence of the essential gene *lpxC* immediately downstream (under the control of its own promoter). Transient integration of the Cam^R marker and subsequent transcription originating from the Cam^R promoter might have had an impact on *lpxC* expression. The genes targeted in the original study were at the distal end of transcriptional units or in single-gene operons. This method may be less successful when the targeted essential gene occurs upstream of other essential genes, though these situations are relatively uncommon.
5. Counterselection with I-*Sce*I was more effective at 30°C than at 37°C. The I-*Sce*I allele used here came from pST98-AS and is missing four amino acids near the N terminus ***(19)***. Other alleles of I-*Sce*I under control of the Ara promoter used in this laboratory work efficiently at 37°C, so the effect may be due to either a small deletion or the level of I-*Sce*I expressed from the *tetA* promoter.
6. As an alternative to the I-SceI counterselection described, an amber mutant can also be made using the positive and negative selection properties of the *tetA* gene from Tn10 ***(20)***. A fusion PCR product was made for *gcpE* using a modified *tetA* in place of the Cam^R gene and flanking I-SceI sites. Unwanted transcription toward the essential gene was reduced by modifications to the *tetR* promoter (which overlaps the *tetA* promoter) and insertion of the T7 te terminator. After integration of this fusion product into the chromosome using RDM/Kan/Ara/Tet plates (10 μg/mL Tet) and subsequent identification of a conditional lethal, cells were grown overnight in RDM/Kan/Ara. The removal of the *tetA* gene was then accomplished by diluting and counterselecting on modified RDM plates containing Kan, 0.2% arabinose, 7 g/L additional NaCl, 72 mM NaH_2PO_4, 0.1 mM $ZnCl_2$, and 12 μg/mL fusaric acid, which is lethal to cells expressing *tetA* ***(20)***. Survivors were then screened for tetracycline resistance. Although this approach worked, the I-SceI system was preferred because it resulted in a tighter selection and a higher percentage of colonies that had lost the drug marker.
7. There have been some reports of mutagenic effects from lambda Red expression ***(21)***, so this method may increase the chances of introducing unwanted secondary mutations. The author has found that the mutagenesis plasmid used in "gene gorging" pACBSR ***(6)*** can be easily removed without inducing lambda Red. Mutants are simply streaked on nonselective medium, and ~25 colonies are then tested for loss of the mutagenesis plasmid by patching. Clones that have lost the mutagenesis plasmid are usually identified this way but in some cases are

not. In those cases, patches are restreaked and rescreened until a plasmid-free clone is found.

Acknowledgments

This method was developed under the supervision of the author's Ph.D. advisor, Frederick R. Blattner, and the basic strategy of introducing and suppressing amber stop codons for functional genomics is his vision. Jeremy Glasner performed early experiments with the suppressor tRNA and provided additional guidance.

References

1. Schmid, M. B., Kapur, N., Isaacson, D. R., Lindroos, P., and Sharpe, C. (1989) Genetic analysis of temperature-sensitive lethal mutants of *Salmonella typhimurium*. *Genetics* **123**, 625–633.
2. Guzman, L. M., Belin, D., Carson, M. J., and Beckwith, J. (1995) Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J. Bacteriol.* **177**, 4121–4130.
3. Judson, N., and Mekalanos, J. J. (2000) TnAraOut, a transposon-based approach to identify and characterize essential bacterial genes. *Nat. Biotechnol.* **18**, 740–745.
4. Forsyth, R. A., Haselbeck, R. J., Ohlsen, K. L., Yamamoto, R. T., Xu, H., Trawick, J. D., et al. (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol. Microbiol.* **43**, 1387–1400.
5. Ji, Y., Zhang, B., Van, S. F., Horn, W. P., Woodnutt, G., Burnham, M. K., and Rosenberg, M. (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science* **293**, 2266–2269.
6. Herring, C. D., Glasner, J. D., and Blattner, F. R. (2003) Gene replacement without selection: regulated suppression of amber mutations in *Escherichia coli*. *Gene* **311**, 153–163.
7. Eggertsson, G., and Soll, D. (1988) Transfer ribonucleic acid-mediated suppression of termination codons in *Escherichia coli*. *Microbiol. Rev.* **52**, 354–374.
8. Normanly, J., Kleina, L. G., Masson, J. M., Abelson, J., and Miller, J. H. (1990) Construction of *Escherichia coli* amber suppressor tRNA genes. III. Determination of tRNA specificity. *J. Mol. Biol.* **213**, 719–726.
9. Herring, C. D., and Blattner, F. R. (2004) Global transcriptional effects of a suppressor tRNA and the inactivation of the regulator *frmR*. *J. Bacteriol.* **186**, 6714–6720.
10. Steege, D. A., and Soll, D. G. (1979) Suppression. In: Goldberger, R. F., ed. *Biological Regulation and Development*, Vol. 1. New York: Plenum Press, pp. 433–485.
11. Edgar, R. S. (1966) Conditional lethals. In: Cairns, G., Stent, G. S., and Watson, J. D., eds. *Phage and the Origins of Molecular Biology*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory of Quantitative Biology, pp. 166–170.
12. Benzer, S., and Champe, S. P. (1962) A change from nonsense to sense in the genetic code. *Proc. Natl. Acad. Sci. U.S.A.* **48**, 1114–1121.
13. Brenner, S., Stretton, A. O., and Kaplan, S. (1965) Genetic code: the "nonsense" triplets for chain termination and their suppression. *Nature* **206**, 994–998.
14. Sarabhai, A. S., Stretton, A. O., Brenner, S., and Bolle, A. (1964) Co-linearity of the gene with the polypeptide chain. *Nature* **201**, 13–17.
15. Engelhardt, D. L., Webster, R. E., Wilhelm, R. C., and Zinder, N. (1965) *In vitro* studies on the mechanism of suppression of a nonsense mutation. *Proc. Natl. Acad. Sci. U.S.A.* **54**, 1791–1797.

16. Capecchi, M. R., and Gussin, G. N. (1965) Suppression in vitro: identification of a serine-sRNA as a “nonsense” suppressor. *Science* **149**, 417–422.
17. Herring, C. D., and Blattner, F. R. (2004) Conditional lethal amber mutations in essential *Escherichia coli* genes. *J. Bacteriol.* **186**, 2673–2681.
18. Neidhardt, F. C., Bloch, P. L., and Smith, D. F. (1974) Culture medium for enterobacteria. *J. Bacteriol.* **119**, 736–747.
19. Posfai, G., Kolisnychenko, V., Bereczki, Z., and Blattner, F. R. (1999) Markerless gene replacement in *Escherichia coli* stimulated by a double-strand break in the chromosome. *Nucleic Acids Res.* **27**, 4409–4415.
20. Bochner, B. R., Huang, H.-C., Schieven, G. L., and Ames, B. N. (1980) Positive selection for loss of tetracycline resistance. *J. Bacteriol.* **143**, 926–933.
21. Murphy, K. C., and Campellone, K. G. (2003) Lambda Red-mediated recombinogenic engineering of enterohemorrhagic and enteropathogenic *E. coli*. *BMC Mol. Biol.* **4**, 11.

22

Statistical Methods for Building Random Transposon Mutagenesis Libraries

Oliver Will

Summary

During the construction of random transposon mutagenesis libraries, four essential statistical issues arise: (1) Computing basic probability results for number of open reading frame knockouts. (2) Estimating the number of new open reading frames that will be knockouts in the next set of clones. (3) Estimating the number of essential open reading frames. (4) Computing the probability that an open reading frame is essential given the distribution of insertions. This chapter examines these issues and evaluates potential solutions using three different approaches: Efron and Thisted's estimator, Will and Jacobs's parametric bootstrap, and Blades and Broman's Gibbs sampler. In doing so, this chapter provides guidance for using the R statistical project to solve these problems.

Key Words: bootstrap; clonal libraries; prokaryote; random mutagenesis; statistics.

1. Introduction

Biologists can economically build large-scale knockout libraries for bacteria by using polymerase chain reaction (PCR), complete genome sequence, and laboratory automation ***(1)***. In the ideal knockout library, each clone contains only one silenced open reading frame (also known as an open reading frame [ORF] knockout), and an ORF knockout clone would exist for every ORF in the genome. Such knockout libraries are critical in determining gene functions and have applications in identifying drug targets, assessing phenotypic variation, and studying gene essentiality ***(2)***.

The approach in Ref. ***1*** and **Chapter 9** to building a large knockout library uses random transposable elements. A transposable element (or a transposon) is a sequence of DNA that can excise itself from one sequence and insert itself into another. Random transposons insert themselves at any target sequence (e.g., any TA site) in a prokaryotic genome. Through manipulating the sequence of a transposon, it can become a powerful tool in constructing a knockout library (**Chapter 2**). If a transposon sequence that includes a stop codon and a PCR primer site lands in an ORF, the resulting protein will be stunted, effectively knocking out the protein's function. The exact location of the

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

transposon can then be located by sequencing at the incorporated primer site across the transposon-chromosome junction and comparing this with genomic data.

Despite advances in technology and the utility of random transposable elements, biologists have yet to make a complete random ORF knockout library for a prokaryote; a class of ORFs for which there will be no knockout clones always exists. A knockout clone for an ORF might not exist because the transposon has not landed in the ORF yet or because the protein encoded by the ORF is essential for cell function. The proteins encoded by essential ORFs are potential drug targets, and random transposon libraries provide a fast method of generating many candidates *(3)*. Identifying essential ORFs has been a crucial undertaking *(4–8)*.

Other chapters in this book concern the laboratory steps needed to create a library. This chapter looks at some of the mathematical questions that arise while building one. Because our concerns are mathematical, we will abstract the construction process. First, we assume that the genome of the prokaryotic organism is sequenced and all the ORFs are annotated. The transposon we use has a target identifiable from the sequence, and the transposon inserts itself with equal probability at each target. For example, Jacobs et al. (*[1]* and **Chapter 9**) used a transposon that putatively inserted with equal probability at each base pair, and Lamichhane et al. *(8)* used a transposon that inserted with equal probability at each TA dinucleotide. Then, we assume that the location of the transposon upon insertion can be accurately determined and that it inserts at exactly one place in the genome. From the assumptions of target identifiability and equal insertion probability and further assuming that there are no essential ORFs, we deduce that the probability of a transposon landing within an open reading frame equals the number of targets in the ORF divided by the number of targets in the genome (**Note 1**). Using these assumptions, we can estimate the following parameters of the construction process:

1. We can compute the expected value and the variance of the number of ORF knockouts in our clonal library provided there were no essential genes. To do this, we use the multinomial model (**Section 3.1**).
2. Given that we found i different ORF knockouts in the first n clones, we can predict the number of new ORF knockouts in the next d clones by using Efron and Thisted's estimator *(9)* (**Section 3.2**) or Will and Jacobs's bootstrap (**Section 3.3**). The predictions by both methods are accurate even if there are essential ORFs present in a genome.
3. Given that we have n clones, we can estimate the number of essential genes m by using Will and Jacobs's bootstrap (**Section 3.3**) or Blades and Broman's Gibbs sampler *(10)* (**Section 3.4**).
4. Given that we have n clones, we can estimate the probability that an ORF is essential using Blades and Broman's Gibbs sampler (**Section 3.4**).

All the estimates from the **Methods** section are summarized in **Table 1**. All methods in this chapter have been programmed in the statistical language R in the *occugene* and *negenes* packages (**Note 2**). We use the following hypothetical example for illustration throughout the chapter. Imagine a circular bacterial genome containing 20 ORFs and 124 transposition target sites, annotated as shown in **Table 2**. Annotating prokaryotic genomes has its subtleties. Generally, bacterial chromosomes are circular and contain open reading frames oriented in both directions. The R package requires target sites to

Table 1
Predictions Generated by Different Strategies for the Expected Number of New Inactivated ORFs to Be Found Within the Next 10 Mutant Strains Tested and for the Total Number of Essential Genes in a Genome

	Number of new ORF knockouts		Number of essential ORFs	
	Simulated	*P. aeruginosa*	Simulated	*P. aeruginosa*
Efron and Thisted	0.11 ± 0.34	0.23 ± 0.02	NA	NA
Will and Jacobs	0.17 ± 0.38	0.18 ± 0.01	6.11 ± 4.94	377 ± 35
Blades and Broman	NA	NA	6.69 ± 1.76	403 ± 24

The first number is the point estimate and the second number is a measure of the spread. NA, not applicable.

be numbered from 1 to *k* in one direction around the genome. For a circular bacterial chromosome, the starting position and the direction of this process are both arbitrary. Table 2 uses the convention of the *occugene* R package and treats the end of the last ORF as the distal-most point of the genome.

Note that all ORFs in the sample genome are located one target apart, except for ORFs 5 and 6, and 15 and 16, which overlap on two targets. In this hypothetical example, 60 independent insertion mutants have been isolated. The coverage (the

Table 2
A Sample Genome Annotation Table

ORF number	ORF start (at target no.)	ORF length (in number of targets)	Essential?	Number of knockouts
1	2	10	No	5
2	13	9	No	3
3	23	8	Yes	0
4	32	7	No	4
5	40	6	No	0
6	44	5	Yes	0
7	50	4	No	2
8	55	3	No	2
9	59	2	Yes	0
10	62	1	No	1
11	64	10	No	7
12	75	9	Yes	0
13	85	8	No	4
14	94	7	No	5
15	102	6	Yes	0
16	106	5	No	0
17	112	4	No	6
18	117	3	Yes	0
19	121	2	No	3
20	124	1	No	1

number of independent mutants divided by the total number of ORFs) of this hypothetical library is 3.0× (60 clones/20 ORFs), which is lower than what one encounters in practice. For example, Jacobs and coworkers have isolated 30,456 mutant strains in the *Pseudomonas aeruginosa* strain PAO1 (*[1]* and **Chapter 9**), the genome of which contains 6,264,403 targets in 5570 ORFs. This corresponds with a random-insertion library with 5.5× coverage. Transposon insertions were recovered in 4909 ORFs, leaving 661 ORFs without hits. In Ref. *1*, the authors reported 678 ORFs missed by transposons. This was due to 17 clones being excluded from analysis for technical reasons. We analyze the *P. aeruginosa* library in this chapter as well.

2. Materials

1. Computer.
2. R statistical package downloadable from http://www.r-project.org.
3. Annotated genome.
4. List of chromosomal locations of recovered transposition events.

3. Methods

The methods below are divided into four sections: (1) the multinomial model, (2) Efron and Thisted's estimator, (3) Will and Jacobs's bootstrap, and (4) Blades and Broman's Gibbs sampler.

3.1. Multinomial Model

One arrives at the multinomial distribution as the model for the number of insertions per open reading frame from the assumptions listed in the introduction. To understand the multinomial distribution, imagine throwing *n* balls into *k* boxes. Where each ball lands is independent of where other balls land, and the probability of landing in the *j*th box is known and expressed as p_j. Formally, the probability of a vector of counts, $(x_1, \dots, x_k)$, is

$$P(x_1, \dots, x_k) = \binom{n}{x_1, \dots, x_k} p_1^{x_1} \dots p_k^{x_k}, \qquad (1)$$

where $\binom{n}{x_1, \dots, x_k} = \dfrac{n!}{x_1! \dots x_k!}$ and is called the multinomial coefficient. The occupancy distribution refers to the number of boxes that have at least one ball.

The multinomial distribution with the probabilities computed from the genome annotation is called the multinomial model. Expected values and variances may be computed exactly or approximately from the multinomial model representing a random mutagenesis library without essential ORFs. This approximation is presented because it is easy to modify and to accommodate other parameters of interest, such as the number of ORFs that will be hit only once.

3.1.1. Exact Computation

Assuming that there are no essential genes and that the assumptions of independence and equal probability of hits (**Section 1**) hold, the numbers of insertions per ORF are distributed as a multinomial random variable with parameters $n, p_1, \dots, p_k$, where *n* is the number of mutant strains generated, p_j is the probability of landing in the jth ORF, and k is the number of ORFs. For simplicity, we ignore the overlaps between ORFs. The probability of landing in an ORF is computed as the number of targets per ORF divided by the total number of targets in the chromosome. If there were overlaps, these probabilities would need to be rescaled to one. For the multinomial model, the number of inactivated ORFs follows the occupancy distribution for a multinomial (*11*). The expected value of the occupancy distribution is

$$E_n = k - \sum_{j=1}^{k} (1-p_j)^n \tag{2}$$

and the variance is

$$k - \sum_{j=1}^{k}(1-p_j)^n - \left[k - \sum_{j=1}^{k}(1-p_j)^n\right]^2 + \sum_{1\le i\ne j\le k}\left[1-(1-p_i)^n-(1-p_j)^n+(1-p_i-p_j)^n\right]. \tag{3}$$

We use E_n to denote the expected number of mutated ORFs within n clones (**Note 3**).

Thus far, we have not accounted for the intergenic regions, the gaps between the ORFs, in our model. We code the gap region as an extra ORF that has as many targets as there are in all the gaps. If we do not include the noncoding region as an ORF, we must multiply n by the fraction of the genome that codes for proteins. Hence, the parameters for the hypothetical example incorporating the intergenic regions are $n = 60$ and $k = 21$. Calculating the sums by hand is tedious, so the R package *occugene* has the functions *eMult* and *varMult* that automate this process. Shown below is how to run the functions on the hypothetical example:

```
> n <- 60
> p <- c(seq(10,1,-1),seq(10,1,-1),18)/124
> p <- p/sum(p)
> eMult(n,p)
[1] 17.74773
> varMult(n,p)
[1] 1.744004
```

The first line in this example assigns the value 60 to the variable n (the number of mutant strains). We compute the probability of a transposon landing in an ORF based on the annotation encoded in the second and third lines. The third line rescales the probabilities to 1 because the 5th and the 6th, and the 15th and 16th ORFs overlap. The 4th and 5th lines invoke the tools that were installed when the *occugene* package was loaded. The function *eMult* returns the expected number of ORFs with at least one insertion, and *varMult* returns the variance.

We expect to recover transposon insertions in about 18 ± 2.6 ORFs within a random library of 60 mutant strains; 12 ORFs with inserts are actually present in our example (**Table 2**). Likewise, within the *Pseudomonas aeruginosa* library (where n = 30,456 and k = 5570), 5275 ORFs (with a variance of 230) were expected to be hit assuming there were no essential ORFs present in the genome. In reality, transposition events were detected in 4909 ORFs (**Note 4**).

3.1.2. Approximation

When n and k become large (of the order thousands), as they are in case of the *P. aeruginosa* random library (*[1]* and **Chapter 9**), the computations of the sums in R become progressively slower. Instead, simulation of a random-insertion library in R can be used to approximate the expected value and variance of the occupancy distribution (**Note 5**). To do so, simulate a multinomial random variable with the probability vector (p_1,K,p_k) above (**Note 6**). Let Y_j be the number of cells with at least one hit in the simulated multinomial, hence, Y_j has the occupancy distribution. Repeat simulations of Y_j l times and use the results to approximate the expected value as

$$\bar{Y} = \frac{1}{l}\sum_{1\le j\le l} Y_j \qquad (4)$$

and the variance as

$$S^2 = \frac{1}{(l-1)}\sum_{1\le j\le l}\left(Y_j - \bar{Y}\right)^2. \qquad (5)$$

In the package *occugene*, the functionalities *eMult* and *varMult* can be used to carry out the **approximations 4** and **5** by including the *iter* parameter in the command line. The *seed* parameter in the example below is for seeding the R random number generator so that the numerical results of the code will be the same as the results in the following example. The commands for approximating the expected value and variance are as follows:

```
> n <- 60
> p <- c(seq(10,1,-1),seq(10,1,-1),18)/124
> p <- p/sum(p)
> eMult(n,p,iter=1000,seed=4)
[1] 17.691
> varMult(n,p,iter=1000,seed=4)
[1] 1.607126
```

Again, n is the number of clones, and vector p contains the probabilities of a transposon landing in a particular ORF or in intergenic regions. This approximation predicts 18 ± 2.5 ORFs to be inactivated in the mock experiment (**Table 1**) and 5272 ORFs (with a variance of 224) within the *P. aeruginosa* library *(1)*.

3.2. Efron and Thisted's Estimator

Efron and Thisted *(9)* estimate the total number of species in existence based on how many times each known species is observed (a few species are spotted only once, others twice, some three times, etc.). These data are similar to what we see in a random knockout library. Transposition events in some ORFs appear once, in others twice, and so on. A formula by Efron and Thisted *(9)* can be used to estimate a number of new inactivated ORFs within the next d clones, given the number of diverse ORF hits in the previous n clones. The estimate is valid as long as $n > d$. The $\Delta(d)$ is a random variable for the number of new knockouts; its mean is estimated as follows:

$$\hat{\Delta}(d) = n_1(d/n) - n_2(d/n)^2 + n_3(d/n)^3 - \ldots \quad \textbf{(6)}$$

and its variance as

$$\text{Var}\ \hat{\Delta}(d) = n_1(d/n)^2 + n_2(d/n)^4 + n_3(d/n)^6 + \ldots, \quad \textbf{(7)}$$

where n_1 is the number of ORFs that appear in the first n clones of a mutant library exactly once, n_2 is the number of ORFs that appear twice, n_3 is the number of ORFs that appear three times, and so forth. The **estimates 6** and **7** are indeed accurate, in spite of the presence of an unknown set of essential ORFs.

3.2.1. Estimating $\hat{\Delta}(d)$ and Var $\hat{\Delta}(d)$ for the Hypothetical Example

To load the example data into R, use functions *loadAnnotation* and *loadExperiment* of the *occugene* package, designed for the input of genomic data and insertion locations from text files. The annotation file must contain four tab-delimited columns labeled as: “idNum,” “first,” “last,” and “orientation.” The Annotation file format for the mock dataset is illustrated in **Table 3**.

The field “idNum” should contain an arbitrary unique identifier. The fields “first” and “last” correspond with the first and last target sequences within an ORF (inclusive) and should appear in the order of targets numbered on a chromosome so that a value in “first” is always smaller than that in the “last.” The field “orientation” features 0 for ORFs oriented from left to right and 1 for ORFs in the opposite direction. The *occugene* package requires that insertion locations are stored in a separate file in a column titled “position.” To load these files into R, type the following string into the R command prompt:

```
> AFILE <- “sampleAnnotation.txt”
> IFILE <- “sampleInsertions.txt”
> a.data <- loadAnnotation(AFILE)
> experiment <- loadInsertions(IFILE)
```

The variables *AFILE* and *IFILE* assume the annotation and insertions file names, respectively. The functionality *loadAnnotation* returns the genome annotation matrix,

Table 3
The *occugene* Package Annotation File (*sampleAnnotation.txt*) Format

idNum	First	Last	Orientation
1	2	11	0
2	13	21	0
3	23	30	0
4	32	38	0
. . .			

and *loadInsertions* does the same for the list of insertions. Alternatively, load the files after the *occugene* package has been installed:

```
> data(sampleAnnotation)
> data(sampleInsertions)
> a.data <- sampleAnnotation
> experiment <- sampleInsertions
```

The *occugene* package contains the hypothetical data set used in this chapter. To gain access to it, use the data function as shown in the first and second lines above. After issuing the command, the data set is accessible through the names *sampleAnnotation* and *sampleInsertions*. The *occugene* package includes the tool *etDelta* that computes Efron and Thisted's estimates, the point estimate of **equation 6** and its **variance 7**. To run the tool to estimate the number of the new inactivated ORFs present in the next 10 clones, type:

```
> orf <- cbind(a.data$first,a.data$last)
> clone <- experiment$position
> etDelta(10,orf,clone)
$expected
[1] 0.1190665

$variance
[1] 0.02936508
```

We combine the first and last columns of the matrix *a.data* to form the two-column matrix *orf* in the first line. The tool *etDelta* requires the two-column format. We assign the position column of the *experiment* matrix to the variable *clone* in the second line in the example input above. In the third line, *etDelta* returns a list with two members. The first member labeled "expected" is the value from **equation 6**, and the second member labeled "variance" is the value from **equation 7**. In the *P. aeruginosa* example, we expect to see 0.23 new ORF knockouts in the next 10 clones with a variance of 7.72×10^{-5}.

The usefulness of Efron and Thisted's estimates is that they can be used as stopping criteria for ending construction. Large random mutagenesis libraries are created in multiwell plates, which usually have a fixed cost to produce. The functions $\hat{\Delta}(d)$ and Var $\hat{\Delta}(d)$ provide a method of monitoring the progress of library construction because at some point the cost of making a new plate will outweigh the gain of the new ORF knockouts generated. Where this point lies is at the discretion of the researcher.

3.3. Will and Jacobs's Bootstrap

Monitoring the confidence interval for the number of essential ORFs provides another method of choosing when to stop library construction. Once the construction of a library has been completed, estimating the number of essential ORFs is important as well. Jacobs et al. (*1*) outline a bootstrap method that estimates the number of essential ORFs and has the flexibility to estimate the number of new ORF knockouts in the next *d* clones, similar to the Efron and Thisted's estimator (**Section 3.2**).

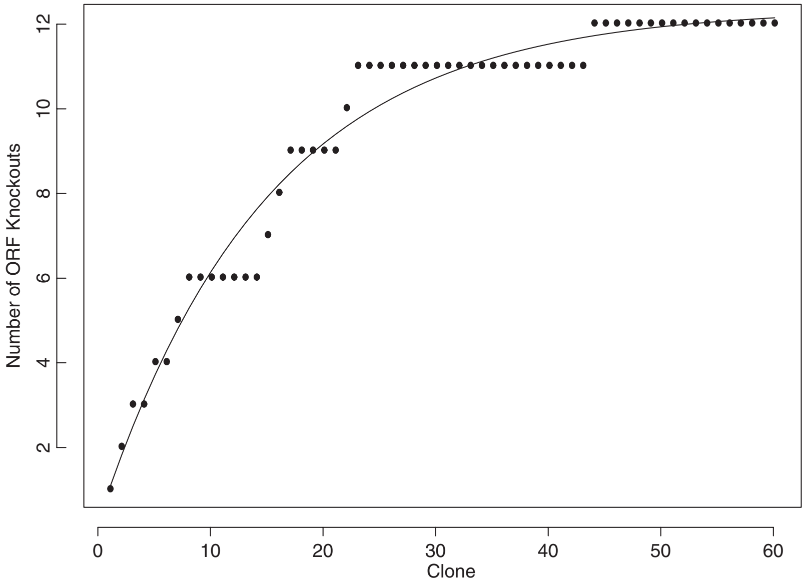


Fig. 1. The dots are the cumulative number of ORFs inactivated by insertions. The line is the best-fitting parameterization, **equation 8**.

To set up the parametric bootstrap, we can fit the function

$$f(n) = b_0 - b_1 \exp(-b_2 n) \tag{8}$$

to the cumulative plot of the number of ORFs hit (**Fig. 1**). The **function 8** is loosely based on the expected value in **equation 2**. The parameters b_0, b_1, and b_2 are chosen to minimize the residual standard sum of squares between the function and the data, and b_0 is a natural estimate of the number of nonessential genes. The value $k - b_0$ is an estimate of the number of essential genes m. For the parameter k, we ignore the artificial ORF created by concatenating the gaps between the actual ORFs. $\Delta(d)$ is estimated with:

$$\hat{\Delta}_0\,(d) = f(n + d) - f(d). \tag{9}$$

As long as the variables *a.data* and *experiment* from **Section 3.2** have been loaded, use the following commands to fit the function in **equation 8** to the hypothetical mutant library:

```
> orf <- cbind(a.data$first,a.data$last)
> clone <- experiment$position
> fFit(orf,clone,FALSE)
Nonlinear regression model
model: noOrfs ~ b0 - b1 * exp(-b2 * x)
data: cumul
   b0    b1    b2
12.34393445 12.05474514 0.06667558
residual sum-of-squares: 15.62024
```

The third value of the *fFit* function, FALSE, in the third line is suppressing the trace of the nonlinear fitting routine. The *fFit* function returns the output from the nonlinear fitting routine. The first line of the output is the name of the routine; the second line is the model specification used in the routine; the third is the internal name for the data set used; the fourth and fifth have the point estimates for b_0, b_1, and b_2; and the sixth has the residual sum of squares from the fitting routine.

The naïve estimates of b_0, b_1, and b_2 are 12.56, 12.20, and 0.06, respectively, which means there are $20 - 12.34 \cong 8$ essential ORFs. The estimates of b_0, b_1, and b_2 are 4893, 4685, and 0.00013 for the *P. aeruginosa* library. There are $5570 - 4893 = 677$ essential ORFs according to the naïve estimates, but there are only $5570 - 4909 = 661$ candidate essential ORFs. We know nothing of the variance and bias of the estimate from **equation 8**, so the fact that the estimated number of essential genes appears larger than the observed number of candidate essential genes in *P. aeruginosa* could be due to bias in the estimator b_0.

Here is how to correct for bias and estimate the variance for the estimate of the number of nonessential genes b_0. Fitting the function in **equation 8** to the cumulative occupancy plot does not have the error structure of a standard nonlinear regression; hence, one must compute the bias and variance of the b_0 differently by using a parametric bootstrap. Our approach is based on Ref. *12*. Proceed by assuming that parameters fitted to the cumulative occupancy distribution with essential genes have the same bias and variance as when they are fitted to a cumulative occupancy distribution of the multinomial model without essential genes. This assumption is written as:

$$b_0 - \hat{b}_0 \sim k - \hat{b}_0^*, \qquad (\mathbf{10})$$

where the * notation indicates values based on the multinomial model without essential genes.

Simulate l multinomial distributions based on the annotation with no essential ORFs. **Section 3.1.1** describes the probability model from which to simulate. Let b_{0j}^* be the b_0 fitted to the j^{th} simulated multinomial. Then, one estimates Bias $b_0 \cong \overline{b_0^*} - k$ where k is the number of ORFs in the genome and $\mathrm{Var}(b_0) \cong 1/(l-1) \sum_{1 \le j \le l} \left(b_{0j}^* - \overline{b_0^*}\right)^2$. The bar notation means to take the mean of the b_{0j}^* as in **equation 4**. The bias-corrected estimate of b_0 is

$$\hat{b}_0 + k - \overline{b_0^*} \qquad (\mathbf{11})$$

and the $(1 - \alpha) \times 100\%$ confidence interval for b_0 is

$$[\hat{b}_0 + k - \hat{b}^*_{0(l[1-\alpha/2])}, \hat{b}_0 + k - \hat{b}^*_{0(l\alpha/2)}], \qquad (\mathbf{12})$$

where $\hat{b}^*_{0(x)}$ is the xth smallest $\hat{b}_0^*$. To find the point estimate and confidence interval for m, the number of essential genes, subtract **equation 11** and **equation 12** from k. To compute the bias-corrected estimator and the confidence interval for the number of nonessential ORFs in the hypothetical data set, we use the *unbiasB0* functionality in the *occugene* package. Enter the following R commands:

```
> orf <- cbind(a.data$first,a.data$last)
> clone <- experiment$position
> unbiasB0(orf,clone,iter=100,seed=4,alpha=0.05)
. . .
$b0
[1] 13.88748

$CI
[1] 8.356225 18.232557
```

The function *unbiasB0* has three additional parameters that can be set as follows. The parameter *iter* is the number of bootstrap replicates, *seed* is the seed for the random number generator, and *alpha* is the type I error rate for the confidence interval. After the third line, the ellipsis indicates that there is trace information omitted from the output. The *unbiasB0* function requires a few seconds to run and returns a list with two members. The first member is the unbiased estimate of b_0, and the second is the confidence interval. The unbiased estimator for the number of essential ORFs in our hypothetical data set is 6, but the confidence interval encompasses most of the range from 2 to 8. For the *P. aeruginosa* library, we find that the unbiased estimate of the number of nonessential ORFs is $b_0 = 5192.754$ with a 95% confidence interval of [5160.088,5229.259]. The number of essential ORFs is around 377 with a 95% confidence interval of [340,410].

We move on to duplicating Efron and Thisted's estimator from **Section 3.2**. Remember the definition of $\hat{\Delta}_0(d)$ from **equation 9**. The naïve estimate is 1.35 for our hypothetical example and 0.11 for the *P. aeruginosa* library using **equation 9** for the number of new ORF knockouts in the next 10 clones. Both were computed using the function *delta0* in the *occugene* package. To estimate the bias and variance in $\hat{\Delta}_0(t)$, we use a parametric bootstrap again in which we assume:

$$E_{n+d} - E_n - \hat{\Delta}_0^*(d) \sim \Delta(d) - \hat{\Delta}_0(d) \qquad \textbf{(13)}$$

where $\hat{\Delta}_0^*(d)$ is the function fit to a data set simulated from the multinomial model in **Section 3.1.1** and the expect value E is computed using **equation 2** (**Note 7**). The unbiased estimator of $\Delta(t)$ becomes:

$$\hat{\Delta}_0 - E_{n+d} + E_n - \overline{\hat{\Delta}_0^*(d)}. \qquad \textbf{(14)}$$

The 100 × (1 − α)% confidence interval is

$$[\hat{\Delta}_0(d) + E_{n+d} - E_n - \hat{\Delta}_0^*(d)_{(l[1-\alpha/2])}, \hat{\Delta}0(d) + E_{n+d} - E_n - \hat{\Delta}_0^*(d)_{(l\alpha/2)}] \qquad \textbf{(15)}$$

in which $\hat{\Delta}_0^*(d)_{(x)}$ is the xth smallest simulated difference. To estimate the number of clones that will have new knockouts in the next 10 clones, use the function *unbiasDelta0:*

```
> orf <- cbind(a.data$first,a.data$last)
> clone <- experiment$position
> unbiasDelta0(10,orf,clone,iter=100,
seed=4,alpha=0.05)
```

```
. . .
$delta0
[1] 0.1745323

$CI
[1] -0.5862741   0.5578337
```

The *unbiasDelta0* tool requires the first three parameters. The first parameter, 10, is the number of new clones that will be made, and *orf* and *clone* come from the first two lines. The last three parameters control the bootstrap. The parameter *iter* is the number of bootstrap simulates, *seed* is for the random number generator, and *alpha* is the type I error. As in other functionalities, *unbiasDelta* returns a two-member list with the first member being the point estimate and the second being the confidence interval. We interpret the –0.586 lower bound of the confidence interval as 0. For the *P. aeruginosa* library, the unbiased estimate for the number of new ORF knockouts in the next 10 clones is 0.18 with a 95% confidence interval of [0.17,0.19]. The parametric bootstrap takes more time to run than Efron and Thisted's estimator.

3.4. Blades and Broman's Gibbs Sampler

Using the same probability model as in **Section 3.1.1**, but with the overlap regions explicitly included (**Note 8**), Blades and Broman ***(10)*** created a Gibbs sampler to estimate the number of essential genes. They programmed the sampler in R and distributed it as the package *negenes* ***(13)***. They have analyzed a random mutagenesis library created for *M. tuberculosis* with their sampler ***(8)***.

We focus on two aspects of the *negenes* package: (a) estimating the number of essential genes as we did with the bootstrap in **Section 3.3** and (b) computing the posterior probability that a gene is essential given the insertion locations already observed. The package uses a different format for the genome annotation table, which is stored as a two-column matrix with the first column labeled "n.sites" and the second column labeled "n.sites2" and rows populated with ID numbers as shown in **Table 4**.

Each row of the matrix corresponds with an ORF. The column labeled "n.sites" lists the number of targets that are located only in one ORF, and the column labeled "n.sites2" has the number of targets shared by this ORF and the next one. The number of targets shown in "n.sites2" for the last ORF is the number of targets shared between this ORF and the first one. The *occugene* package has the function *occup2Negenes* to convert the annotation table and the list of insertions used in *occugene:*

Table 4
The *negenes* Package Annotation Matrix Format

	n.sites	n.sites2
1	10	0
2	9	0
3	8	0
4	7	0
. . .		

```
> orf <- cbind(a.data$first,a.data$last)
> clone <- experiment$position
> newOrf <- occup2Negenes(orf,clone)
> newOrf
      n.sites  n.sites2   counts  counts2
[1,]       10         0        5        0
[2,]        9         0        3        0
. . .
```

The function *occup2Negenes* combines the list of insertions and the genome annotation into one matrix. The column "counts" contains the number of insertions that are recovered in each ORF, and "counts2" has the number of insertions that are recovered in the overlapping region. One is ready to run the Gibbs sampler after using the conversion function. To do so, type the command (**Note 9**):

```
> output <- negenes(newOrf[,1],newOrf[,3],
                    newOrf[,2],newOrf[,4])
```

The output object has three fields once the sampler has completed its computations: "n.essential," "summary," and "geneprob." The first field, "n.essential," has the posterior sample for the number of essential ORFs from the Gibbs sampler run. The second field, "summary," has the estimate for the number of essential genes. For the hypothetical example, it is:

```
> output$summary
mean            sd             2.5%          97.5%
6.6891596       0.8768055      5.4750000     8.0000000
```

which states that there are about seven essential genes with a 95% credible interval of [5,9]. The *mean* value is the mean of the posterior distribution, *sd* is the standard deviation, 2.5% is the lower bound, and 97.5% is the upper bound. The final field, "geneprob," has the posterior probabilities that a gene is essential. Recall that the hypothetical example is created with ORFs 3, 6, 9, 12, 15, and 18 being essential. The sampler gives the probabilities of these ORFs being essential as 0.98, 0.78, 0.65, 0.99, 0.86, and 0.77. The estimated number of essential ORFs for the *P. aeruginosa* library is 403 essential genes with a 95% credible interval of [381,428]. **Table 1** summarizes all the results from the chapter (**Note 10**).

Notes

1. Biologists find it easier to annotate prokaryotic genomes than eukaryotic ones. Tables such as **Table 2** are commonly generated once an entire genome has been sequenced. There is one problem with deriving annotations in this manner; namely, it is unclear how much of the protein is needed to maintain function. So insertions in the distal region of the gene, close to the 3′ end, stunt the gene but might not completely impair function. Sometimes, targets that are close to the 3′ end are not counted as sites that eliminate protein function. The *P. aeruginosa* annotation has not been edited by ignoring base pairs close to the 3′ end.

2. In general, the most difficult part of the procedures in the chapter is using the R project for statistical computing (*14*). R is the open-source version of S-PLUS, and it contains a large number of routines written by leading statisticians. Both *occugene* and *negenes* packages must be installed into R, and then the user manipulates them in the R environment. The installation procedure for the packages varies depending on the computer platform. One mainly uses a command line to interact with R, and a few people find it difficult. Unfortunately, this chapter cannot serve as an introduction to R; however, you can download R free of charge from http://www.r-project.org/, where ample documentation and directions for accessing helpful mailing lists are also available. There are also numerous books on how to use R, and Krause and Olson's book (*15*) is especially useful for beginners.
3. The expected value and variance formulas in **Section 3.1.1** are derived from analyzing the random variable that counts the number of ORFs missed. Writing the random variable as a sum of indicator functions shows us how the expressions were derived. If X_j is the number of different clones in which ORF j has a transposon insertion, the random variable for the number of ORFs missed is

   $$I(X_1 = 0) + \ldots + I(X_k = 0). \qquad (\mathbf{16})$$

   The expected value is computed using relation:

   $$\mathrm{E}\left[\sum I(X_j = 0)\right] = \sum \mathrm{E}I(X_j = 0). \qquad (\mathbf{17})$$

   Likewise, the variance becomes:

   $$\mathrm{Var}\left[\sum I(X_j = 0)\right] = \mathrm{E}\sum I(X_j = 0)^2 + \mathrm{E}\sum_{i \neq j} I(X_i = 0) I(X_j = 0) - \left(\mathrm{E}\left[\sum I(X_j = 0)\right]\right)^2 \qquad (\mathbf{18})$$

   From these sums and the fact that $\mathrm{E}I(X_j = 0) = P(X_j = 0)$, we derive the formulas in **Section 3.1.1**. For a more in-depth explanation, see Ref. *11*.
4. We can use the multinomial model (**Section 3.1.1**) to predict the size of a random library required to generate at least one insertion in every ORF. Assuming that the probabilities of a transposon landing in an ORF are equal for all ORFs, the required number of independent mutants can be computed as $k \log k$ (base e). For the hypothetical example in this chapter, it amounts to $20 \log 20 \cong 60$, and for the *P. aeruginosa* PAO1 library ([*1*] and **Chapter 9**) it is $5570 \log 5570 \cong 48042$.
5. It may appear illogical to present the exact computation of the expected value and variance in **Section 3.1.1** and to follow that with mere approximation in **Section 3.1.2**. The addition of **Section 3.1.2** was provoked by practical questions that arose during the *P. aeruginosa* library construction, such as: "How many ORFs on average will be inactivated only once in the entire library (a number of unique mutant strains represented by a single clone)?" or "How many ORFs devoid of inserts are expected to appear next to an ORF that has been hit exactly once by chance alone?" Theoretically, formulas similar to **equation 1** and **equation 2** can be derived to answer each question, but in practice simulations are much easier to perform.
6. R includes many tools for simulating the multinomial distribution (**Section 3.1.2**); however, they are not included in a standard package. For example, the package *combinat* contains a simulation tool. The package *negenes* includes the *sim.mutants* tool capable of simulating a random-insertion library with potentially overlapping and/or essential ORFs. The package *occugene* features a multinomial simulator as well but without direct user access.

7. The R functions *fFit* and *unbiasB0* accommodate overlapping ORFs. If a single transposon lands in two ORFs due to hitting a shared target, two knockouts are counted. However, the bootstrap for the number of the new knockouts in the next *d* clones does not accommodate overlapping ORFs. In computing the expected values for the occupancy distribution, the function *unbiasDelta0* computes the probability of insertion as the length of the ORF divided by the total length of the genome. The probabilities are normalized to 1 if needed. Hopefully, the overlap of the ORFs in the genome is small enough so that the equations in **Section 3.1.1** do approximate the number of ORF knockouts.
8. To include the overlapping ORFs into the model, Blades and Broman ***(10)*** use an extended model based on the one in **Section 3.1.1**. Recall that *k* is the number of ORFs in the chromosome and *n* is the number of mutants created. Let *m* be the number of essential genes. Some ORFs overlap slightly in bacterial genomes. Now, let x_j be the number of clones that have an insertion solely in the *j* th ORF. Let y_j be the number of clones that have an insertion in the shared region between ORF *j* and *j* + 1. It is convenient to use x_k as the number of clones with insertions in the intergenic region and to renumber the ORFs appropriately. Let $(p_1, \ldots, p_k)$ be the probabilities that a transposon lands in one of the target sequences located exclusively in one ORF, and let $(q_1, \ldots, q_k)$ be the probabilities that a transposon lands in one of the shared targets. Usually, the probabilities are computed as the number of targets in the region, divided by the total number of targets in the chromosome. Let the unobserved vector $G = (g_1, \ldots, g_k)$ of ones and zeros indicate whether an ORF is nonessential or not; 1 means nonessential and 0 means essential. Note that $\sum x_j + y_j = n$, $\sum g_j = k - m$, and $\sum p_j + q_j = 1$. Given an instance of *G* the distribution of (X,Y) is

$$P(X,Y|G) = \binom{n}{x_1, y_1, \ldots, x_k, y_k} \frac{\prod_{1\le j\le k} (p_j g_j)^{x_j} (q_j g_j g_{j+1})^{y_j}}{\left(\sum_{1\le j\le k} p_j g_j + q_j g_j g_{j+1} \right)^n}. \quad \textbf{(19)}$$

The Gibbs sampler uses a prior distribution on *G* where each value of $m \in [1,k]$ is equally likely and, given *m*, each *G* where $\sum g_j = k - m$ is equally likely. Blades and Broman's proposal distribution is described in their technical report ***(10)***.
9. Gibbs samplers are tricky to run properly. There are issues with burn-in and convergence that are hard to assess. The function *negenes* in **Section 3.4** has controls described in the documentation for running the sample. All the nuances of Markov chain monte carlo (MCMC) estimation cannot be explained in this short chapter. However, it has been my experience that the Blades and Broman's Gibbs sampler mixes very well using the default parameters.
10. Throughout **Section 2**, we did not compare the methods. The number of new ORF knockouts can be estimated with either Efron and Thisted's estimator in **Section 3.2** or Will and Jacobs's bootstrap in **Section 3.3**, and the number of essential ORFs can be estimated with either Will and Jacobs's bootstrap in **Section 3.2** or Blades and Broman's Gibbs sampler in **Section 3.4**. The methods have not been subjected to a detailed comparison, but after using all three on real and simulated data, the author recommends Efron and Thisted's estimator for the number of new ORF knockouts and Blades and Broman's Gibbs sampler for the number of essential ORFs (*see* ***[16]***). The bootstrap is noticeably slower than the other two methods, and sometimes the procedure for fitting the parametric **function 8** to the occupancy distribution does not converge, especially at low coverage.

Acknowledgments

I thank the Allan Wilson Centre for supporting me as a postdoctoral fellow. I thank Michael A. Jacobs for helpful discussion about this chapter and Mike Steel and Margee Will for editorial advice.

References

1. Jacobs, M. A., Alwood, A., Thaipisuttikul, I., Spencer, D., Haugen, E., Ernst, S., et al. (2003) Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 14339–14344.
2. Hayes, F. (2003) Transposon-based strategies for microbial functional genomics and proteomics. *Annu. Rev. Genet.* **37**, 3–29.
3. Fraser, C. M. (2004) A genomics-based approach to biodefence preparedness. *Nature Rev. Genet.* **5**, 23–33.
4. Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balázsi, G., Ravasz, E., Daugherty, M. D., et al. (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684.
5. Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., et al. (2002) Functional profiling of *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391.
6. Hutchison IC. A. III, Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., et al. (1999) Global transposon mutagenesis and a minimal mycoplasma genome. *Science*. **286**, 2165–2169.
7. Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., Arnaud, M., et al. Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4678–4683.
8. Lamichhane, G., Zignol, M., Blades, N. J., Geiman, D. E., Dougherty, A., Grosset, J., et al. (2003) A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 7213–7218.
9. Efron, B., and Thisted, R. (1976) Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*. **63**, 435–447.
10. Blades, N. J., and Broman, K. W. (2002) Estimating the number of essential genes in a genome by random tranposon mutagenesis. *Technical Report MS02-20*. Baltimore, MD: Johns Hopkins University.
11. Johnson, N. L., and Kotz, S. (1977) *Urn Models and Their Application: An Approach to Modern Discrete Probability Theory*. New York: John Wiley & Sons.
12. Davison, A. C., and Hinkley, D. V. (1997) *Bootstrap Methods and their Applications*. Cambridge: Cambridge University Press.
13. Broman, K. W. (2004*) negenes: Estimating the Number of Essential Genes in a Genome*. R package version 0.98-5. Available at http://www.biostat.jhsph.edu/~kbroman/software/negenes.html.
14. R Development Core Team (2005) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at http://www.R-project.org.
15. Krause, A., and Olson, M. (2000) *The Basics of S-PLUS*. New York: Springer-Verlag.
16. Will, O., and Jacobs, M. A. (2006) Estimating the number of essential genes in random transposon mutagenesis libraries. Available at http://arxiv.org/abs/q-bio.OT/0608005.

23

Statistical Evaluation of Genetic Footprinting Data

Gábor Balázsi

Summary

As transposomics is extended to genome scale, appropriate statistical methods need to be developed to assign significance to gene essentiality. In this chapter, the author presents a set of steps that, together with genome-scale insertion data and the complete genome sequence of a prokaryote, can be used to classify the genes of the organism as either "essential" or "nonessential."

Key Words: essentiality; genetics; insertion; mutagenesis; Poisson distribution; significance; transposomics.

1. Introduction

The number of genes in prokaryotes can reach a few thousand *(1–3)*, but many of these genes are dispensable. Identifying the genes that are essential in various conditions can result in a better understanding of prokaryotic biology, a better functional annotation of gene products, and the development of more efficient antibiotics.

One of the genome-wide gene essentiality screens used a Tn*5*-based transposome mutagenesis system and identified 620 essential genes and 3126 nonessential genes in *Escherichia coli* (*[4]* and **Chapter 6**). With the extension of transposomics to genome scale, it becomes crucial to develop statistical methods to reliably identify essential genes and assign significance to essentiality calls.

A statistical approach to transposomics is presented in the next section. This approach assumes that insertions are random events that resemble a Poisson process over large portions of the chromosome. The author discusses two biological factors that influence the validity of this assumption: variation of insertion density along the chromosome and the contribution of essential genes to reduce the number of insertions. The possible pitfalls of the technique are discussed briefly at the end of the chapter.

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

2. Materials

In addition to a workstation that can be programmed in a programming language such as C, Perl, or Java, the following data are needed to identify the essential genes of a prokaryote:

1. Transposon insertion locations for the whole genome.
2. A completed genomic sequence of the prokaryote.
3. The most complete annotation of all open reading frames (ORFs) in the genome.

3. Methods

The basic assumption of transposon mutagenesis is that trasposon insertions occur randomly and with uniform density throughout the chromosome. After mapping the insertions along the chromosome, genes without insertions are likely candidates to be essential. However, genes can also be missed by chance, and labeling all genes without insertions as "essential" will generate many false positives. It is therefore necessary to reduce the number of false positives by assigning significance to genes with no insertions.

Intuition tells us that if a gene is very short, or if the insertion density is very low, the gene can easily be missed by insertions. In general, if the insertion density is *r*, the probability of *N* insertions occurring within a DNA region of length *L* is given by the Poisson distribution *(5)*:

$$P_N(L) = \frac{(rL)^N}{N!} e^{-rL}, \tag{1}$$

and therefore, the probability to have no insertions in a gene of length *L* (measured in base pairs) is

$$P_0(L) = e^{-rL}. \tag{2}$$

If the insertion density *r* were known, this formula could be used to determine the significance of essentiality calls. However, *r* is unknown, and therefore it has to be determined prior to the classification of genes according to their essentiality.

The simplest way to determine the insertion density *r* might be to divide the total number of insertions N_T mapped around the chromosome by the length of the full chromosome, L_T:

$$r = \frac{N_T}{L_T}. \tag{3}$$

However, this simplistic approach could be misleading for two reasons. First, nothing guarantees that the insertion density along the chromosome is constant (**Note 1** and **Fig. 1A**). Second, since essential genes on the chromosome exclude insertions, **equation 3** will underestimate the insertion density (**Note 2** and **Fig. 1A**).

To avoid the first problem (variation of insertion density along the chromosome), *r* should be estimated locally instead of globally. To estimate *r* locally, the number of insertions should be determined within a DNA region surrounding the gene, rather than the whole chromosome. To avoid the second problem (the bias introduced by essential

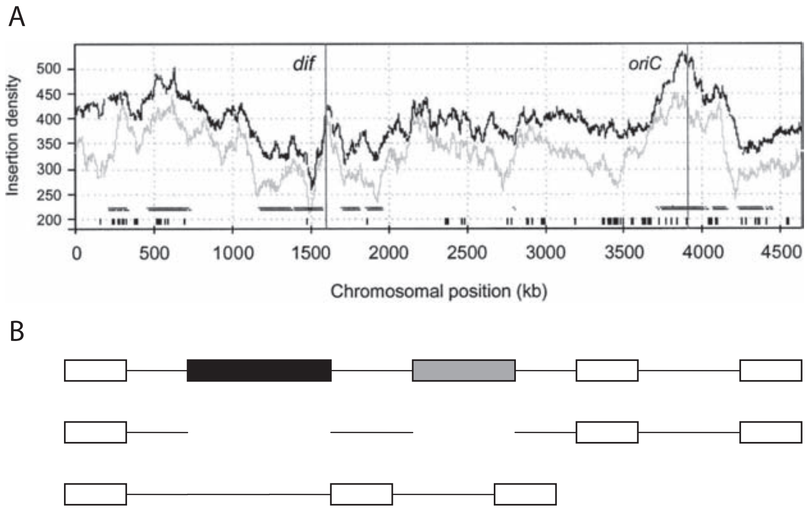


Fig. 1. **(A)** Distribution of transposon insertion densities along the *E. coli* chromosome. Gray lines show the transposon insertion densities calculated as the number of transposition events per 100-kb sliding window over the entire *E. coli* MG1655 chromosome. Values indicated by the black lines were computed in a similar manner, except that all chromosomal regions corresponding with essential and ambiguous genes were excluded from the calculations in order to reconstruct insert distribution prior to selective outgrowth. Gaps in the data (chromosomal regions where transposition events could not be detected due to technical reasons) are indicated by short vertical lines along the *x* axis. The regions where the distributions of transposition events significantly deviate ($p < 0.01$) from a Poisson process are marked by horizontal double lines. *OriC* shows the origin of chromosomal replication, and *dif* denotes the *dif* locus within the replication termination area. (Reprinted from Ref. *4* with permission from American Society for Microbiology.) **(B)** Correcting the bias introduced by essential genes. For the estimation of transposon insertion density within a DNA region, genes with no insertions (or, ideally, all known ORFs) should be left out from the analysis to eliminate the bias of essential genes, which exclude insertions. Shading indicates nonessential genes (white), essential gene (black), and gene with no insertions—a new candidate for essentiality (gray).

genes), the insertion density should be determined only within noncoding regions along the chromosome. This will ensure that essential genes will be excluded and will not cause a bias in the insertion density (**Notes 3** and **4**).

How long should the chromosome region be for a reliable local estimation of the insertion density? Insertion density is estimated by counting the number N of insertions and dividing it by the length L of the DNA in which they occur:

$$r_{est} = \frac{N}{L}. \tag{4}$$

As one would expect, the average of r_{est} is

$$\langle r_{est} \rangle = \frac{\langle N \rangle}{L} = \frac{1}{L}\sum_{N=0}^{\infty} NP_N(L) = r. \tag{5}$$

However, even if the rate of insertions is constant along the chromosome, the number of insertions in DNA segments of identical length L will fluctuate around rL because of the random nature of insertions events. As a consequence, there will be an error in determining r_{est}. The magnitude of this error can be measured by the variance:

$$\langle r_{est}^2 \rangle - \langle r_{est} \rangle^2 = \frac{\langle N^2 \rangle - \langle N \rangle^2}{L^2} = \frac{r}{L}. \qquad \textbf{(6)}$$

According to this formula, the error committed in the estimation of r is higher for short DNA regions. Therefore, the DNA region should be as long as possible without being influenced by regional fluctuations of the insertion density along the chromosome (**Note 4**).

To summarize, for a proper assessment of gene essentiality, the following steps should be taken:

1. Select a gene with no insertions.
2. Exclude all the known ORFs from the DNA (or all genes with no insertions) surrounding the gene to minimize the bias introduced by essential genes, which reduce insertion density (**Note 4** and **Fig. 1B**).
3. Paste together the DNA fragments remaining after the exclusion of all coding regions until the desired length L is reached. The region used to determine the local density should be as long as possible without being affected by fluctuations of insertion density along the chromosome.
4. Using the noncoding DNA, determine the local density of insertions around the gene.
5. Use **formula 2** and the local insertion density r to determine the probability for the gene to be missed by chance alone.
6. Establish a cutoff (**Note 5**). If $P_0(L) < c$ (the probability of being missed by chance is below the cutoff) label the gene as "essential." Otherwise, label the gene as "nonessential."
7. Repeat **steps 1** to **4** for all genes and for various values of L and c (**Note 5**).

Notes

1. DNA replication is a known factor that could result in a location-dependent insertion density. In exponential growth, bacteria are known to initiate a new round of replication before the previous round has terminated *(6)*. Therefore, it is possible to have 2, 4, 8, or even 16 copies of the origin of replication compared with the terminus. As a result, a higher amount of DNA is available for insertion around the origin, and therefore insertion density is expected to be highest around the origin and decreasing toward the terminus. This has indeed been observed in the genome-scale footprinting study (*[4]* and **Chapter 6**).
2. Comparing the insertion density along the *E. coli* chromosome with the insertion-free coding regions included and excluded reveals that r is higher for the latter throughout the chromosome *(4)*. The difference between the two estimates of the insertion density is highest near the origin and lowest near the terminus, which could be explained by the higher density of essential genes near the origin of replication *(7, 8)*.
3. The percentage of coding DNA is much higher in prokaryotes than in higher organisms, and therefore excluding all known ORFs from the DNA might reduce the remaining amount of DNA too much and might lead to poor statistics. An alternative could be to exclude only the ORFs with no insertions from the DNA, but this could artificially increase the local insertion density.

4. The density of genes in some chromosomal regions is higher. In this case, by excluding the coding regions and pasting together the noncoding DNA, the distance from the assessed gene might increase too much. To avoid this problem, a critical distance could be established that cannot be exceeded when estimating insertion density around a gene. This will also result in a maximum limit of L, the number of base pairs used for the estimation.
5. The value of the cutoff c used to classify genes as "essential" or "nonessential" and the length of the DNA region used to determine the insertion density are somewhat arbitrary. Essentiality calls should be confirmed by alternate experimental methods to find the optimal value of c and L. Typically, $L = 10{,}000$ base pairs and $c = 0.01$ are acceptable values to start the analysis.

References

1. Deng, W., Burland, V., Plunkett, G. 3rd, Boutin, A., Mayhew, G. F., Liss, P., et al. (2002) Genome sequence of *Yersinia pestis* KIM. *J. Bacteriol.* **184**, 4601–4611.
2. Tettelin, H., Nelson, K. E., Paulsen, I. T., Eisen, J. A., Read, T. D., Peterson, S., et al. (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae. Science* **293**, 498–506.
3. Blattner, F. R., Plunkett, G. 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., et al. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474.
4. Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balázsi, G., Ravasz, E., Daugherty, M. D., et al. (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684.
5. Harris, J. W., and Stocker, H. (1998) *Handbook of Mathematics and Computational Science*, 1st ed., New York: Springer-Verlag.
6. Donachie, W. D. (1968) Relationship between cell size and time of initiation of DNA replication. *Nature* **219**, 1077–1079.
7. Rocha, E. P. (2004) The replication-related organization of bacterial genomes. *Microbiology* **150**, 1609–1627.
8. Rocha, E. P. (2004) Order and disorder in bacterial genomes. *Curr. Opin. Microbiol.* **7**, 519–527.

24

Modeling Competitive Outgrowth of Mutant Populations: Why Do Essentiality Screens Yield Divergent Results?

Alexander I. Grenov and Svetlana Y. Gerdes

Summary

Mutant propagation (outgrowth) is an important step in all large-scale gene essentiality experiments, profoundly influencing essentiality assignment produced. Using a simplified mathematical model of competitive outgrowth in a diverse mutant population, we have identified several technolgical factors (duration of outgrowth, sensitivity of the scoring technique, initial cell titer of each mutant in the population) that have the largest impact on the outcome of the essentiality screen. The model can be used for planning a large-scale gene essentiality screen as well as for analyzing its results, including meaningful comparisons of "essential" gene lists generated by different techniques.

Key Words: batch culture; microbial growth; transposition.

1. Introduction

Experimental approaches to identification of genome-wide gene essentiality in bacteria include construction of comprehensive genome-wide mutant collections, various techniques utilizing high-throughput random transposon mutagenesis, gene inactivation via inducible expression of antisense RNAs, and other methods, many of which are described in this volume. It is often noted, however, that the lists of essential genes reported by different genome-scale techniques in the same organism under similar growth conditions are often not congruent ***(1–4)***. Even the percentage of genes asserted as "essential" can vary greatly. A number of potential causes for this discrepancy have been identified ***(1–4)***, including transposition site bias, accumulation of secondary transposition events or secondary mutations potentially complementing the original one, potential polar effects of knockout constructs on expression of downstream genes, differences in medium composition, aeration levels, cell densities, use of different techniques for scoring essential versus nonessential genes, and so on.

Although the existing experimental strategies differ in these and many other technical details, they invariably include three stages: (1) generation of a collection of mutants, (2) propagation (outgrowth) of this collection, and (3) assessment of mutant

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

viability. We argue that the main distinction between techniques, profoundly influencing gene essentiality assignments, is whether the outgrowth of each mutant (stage 2) occurs clonally (as in systematic collections of targeted mutants (e.g., *see* **Chapters 9** to **12**) or in a mixed population (as in the majority of random mutagenesis transposition–based techniques, (e.g., **Chapters 3** to **6**). In both cases, gene "essentiality" is deduced from the inability of a cell harboring mutation in a specific gene to undergo a required number of cell divisions necessary to pass the detection threshold of a technique. However, the threshold is much more stringent in a diverse population than in clonal studies. A mutant with a decreased growth rate will be quickly selected against in a population in a planktonic culture, whereas it might be capable of forming a viable colony (albeit slowly) on an agarized medium. This will cause a gene to be scored as "essential" in populational but "dispensable" in clonal essentiality screens (under otherwise identical growth conditions). Furthermore, it is intuitively clear that two technically similar populational essentiality studies can yield significantly different results if they differ in the duration of outgrowth and/or sensitivity of the readout. Here we have attempted to simplistically model the process of competitive outgrowth in a diverse mutant population in order to (a) identify those technological factors that have the largest impact on the essentiality assignments produced by large-scale essentiality screens; (b) determine the minimal number of population doublings (duration of outgrowth) necessary and sufficient for detection of the entire compliment of "essential" genes with the minimal number of false-negative and false-positive assertions; and (c) estimate the specific growth rate μ of a mutant in a mix given the time of its disappearance from the population.

2. Materials

1. Computer.
2. Large-scale essentiality experiment documentation.
3. MATLAB software package (The MathWorks, Natick, MA).

3. Methods

Over the past three decades, the mathematical theory of microbial competition has been a subject of intense investigation. A detailed mathematical description of competition in a chemostat has been reviewed, for example, in Ref. *5* (and references therein). However, application of the existing complex mathematical models to mixed bacterial cultures incorporating more than two competing bacterial species is rare ***(6, 7)*** and would be practically impossible for a mix of 10^4 to 10^5 diverse strains, commonly produced and cocultivated in global gene essentiality screens. We suggest here a simplified model that roughly estimates the role of different experimental factors in generating gene essentiality assignments. The model can be used for planning a large-scale populational gene essentiality screen as well as for analyzing its results, including comparison of "essential" gene lists obtained by different techniques.

3.1. Experimental System

We use the genome-wide gene essentiality experiment described in Ref. *8* and **Chapter 6** as an example throughout this chapter. Briefly, a library of 2×10^5 independent *E. coli* insertion mutants was generated and propagated as a mix in a 1-L fermenter in enriched Luria-Bertani broth for 12 h (approximately 23 population doublings) in batch culture. Cell culture was allowed to reach a late logarithmic growth stage with cell density of 1.4×10^9, at which point genomic DNA was isolated and used to generate genetic footprints. In this experiment, 620 (14%) *E. coli* genes were identified as essential and 3126 (73%) as dispensable for robust aerobic growth in rich media (*[8]* and Chapter 6).

3.2. Basic Assumptions of the Growth Model

In our mixed cell population model, we consider a boundless exponential growth function for each mutant cell type rather than a bounded Gompertz function *(9)* or more complex sigmodal functions *(10)* (**Note 1**). The latter functions can be used to improve accuracy of the numerical solutions in some cases; however, it would be difficult, even impossible, to derive analytical formulas.

At time t, let $q(t)$ denote the number of cells in the population. We use a simple hypothesis that the rate of change of $y = q(t)$ with respect to t is directly proportional to y, or using a differential equation, it can be written as:

$$\frac{dy}{dt} = \mu y, \tag{1}$$

where μ is a specific mutant growth rate, which can be interpreted as a slope from the linear part of the logarithmically transformed cell numbers over time (growth curve). It is related to the mutant doubling time (T_{dbl}) as:

$$\mu = \ln 2/T_{dbl}. \tag{2}$$

The solution of **equation 1** gives:

$$y(t) = y_0 e^{\mu t},$$

where y_0 is the initial number of cells in a fermenter (the inoculate). In practice, we have multiple mutant cells, each growing at its specific growth rate μ_i starting with initial number of cells y_{0i}. Then, the logarithmic phase of growth of a mixed population of N cell types in batch culture can be expressed as the following function Y over time:

$$Y(t) = \sum_{i=1}^{N} y_{0i} e^{\mu_i t}. \tag{3}$$

In our simplified model, we consider two classes of cells: one class, K (for “known”), includes all the mutants in which transposon insertions did not result in a measurable growth rate reduction. This is the larger class of L different species ($L < N$) in which all mutants grow at or near wild-type average rate β_K. The second class, U (for “unknown”), is the mix of the other $N - L$ mutants growing with an average growth rate β_U. In all the mutants in U class, a transposition event has resulted in significant growth rate reduction compared with that of the wild type. We require that the slowest factor $\mu_K^{\min}$ in the first class is significantly greater than the fastest factor $\mu_U^{\max}$ in the second class of mutants, that is,

$$\mu_K^{\min}/\mu_U^{\max} > R, \quad R > 1.5. \tag{4}$$

These requirements are essential to distinguish the slow-growing class of mutants from the *K* class within a diverse population.

3.3. Competitive Outgrowth of a Diverse Mutant Population in Batch Culture

For a mixed population of two classes with different average growth rates β_K and β_U, the cumulative number of cells in a closed batch culture during the *log* growth phase can be approximated as the following growth function of time *t*:

$$Y(t) = Y_K(t) + Y_U(t) = y_{0K}e^{\beta_K+} + y_{0U}e^{\beta_U+}, \qquad (\mathbf{5})$$

where y_{0K} and y_{0U} are initial numbers of cells in known and unknown classes, respectively.

If we are interested in estimating the specific growth rate of a particular mutant *x*, **approximation 5** can be modified to:

$$Y(t) = Y_K(t) + Y_{U^-}(t) + Y_x(t) = y_{0K}e^{\beta_K+} + y_{0_{U^-}}e^{\beta_{U^-}+} + y_{0x}e^{\beta_x+}, \qquad (\mathbf{6})$$

where a third term comprising a single mutant has been added, and class U^- represents the "unknown" class without the mutant *x*.

Let ε denote a sensitivity threshold of a specific detection method employed in a particular essentiality screen. For any given technique, a sensitivity threshold can be expressed as the minimal concentration of a specific cell type in the total population at which this mutant can still be reliably detected (**Note 2**). For a mutant *x*, its relative concentration γ_x in the mix is

$$\gamma_x = Y_x(t)/Y_\Sigma(t). \qquad (\mathbf{7})$$

Using **equation 6**, we can expand **equation 7** as:

$$\gamma_x = \frac{y_{0x}e^{\beta_x+}}{Y_\Sigma(t)} = \frac{y_{0x}e^{\beta_x+}}{y_{0K}e^{\beta_K+} + y_{0U^-}e^{\beta_{U^-}+} + y_{ox}e^{\beta_x+}}. \qquad (\mathbf{8})$$

Because $\mu_K/\mu_U \gg R$ (**condition 4**), and the initial number of mutants in class *K* is comparable with the initial number of mutants in class *U*, then it is clear from **equation 8** that γ_x approaches 0 over time and reaches the sensitivity threshold ε at the time T_x. Hence, the concentration of mutant *x* in the population will reach the minimal level detectable by the sensor and drop below it, in spite of the fact that the population as a whole is expanding logarithmically. In the example experiment (**Section 3.1**), a mutant was scored as absent if its concentration dropped below 1 cell in 100,000. From **equation 8**, we can estimate the specific growth rate of the mutant *x* as follows:

$$\beta_x = \frac{1}{T_x}(\ln(Y_\Sigma) + \ln(\varepsilon/y_{ox})), \qquad (\mathbf{9})$$

where Y_Σ is the total number of cells in the complex population at the time T_x and can be measured experimentally. Conversely, given the mutant's maximum specific growth rate, we can predict the time $t = T_x$ when it will become undetectable by the sensor by swapping β_x and T_x locations in **approximation 9**.

Approximation 9 can be used for planning a large-scale gene essentiality screen as well as for analyzing its results, including the comparison of "essential" gene lists obtained by different techniques. It is clear that (a) the sensitivity ε of the technique used in scoring, (b) the duration of the outgrowth, and (c) the initial cell titer of each mutant largely determine which mutations will be scored as "lethal" (and hence, the corresponding genes will be asserted as "essential"). The longer the propagation time and the higher the detection threshold, the more mutants will appear undetectable, and the more genes will be asserted as "essential." Conversely, a shorter outgrowth and/or a more sensitive screening technique will result in a smaller number of genes scored as essential.

In practical terms, **approximation 9** assigns a numerical value (in terms of growth rate reduction) to the fraction of mutants scored as "essential" in each specific global gene essentiality study and allows meaningful comparison of the data sets produced in different experiments. This is illustrated with the example below.

3.4. A Case Study

In the gene essentiality study reported in Ref. *8* and **Chapter 6**, the average number of cells of each type (with mutations in the same ORF) at the time $t = 0$ (just prior to outgrowth) was 22 (**Note 3**). A gene was scored as "essential" if no transposon insertions could be detected within an ORF after the 12-h outgrowth. The sensitivity threshold ε of the readout procedure (nested PCR) was 10^{-5}. It follows from **approximation 9** that mutants scored in this experiment as "lethal" were growing at or below the following growth rate:

$$\beta_{x1} = \frac{1}{12}\left(\ln\left(1.4\times10^{9}\right) + \ln\left(10^{-5}/46\right)\right) = 0.477. \qquad (\mathbf{10})$$

This "cut-off" growth rate of 0.538 corresponds with the doubling time of 1.29 hours (**equation 2**). This amounts to ~40% of the growth rate of the wild type ($T_{dbl} = 0.52$ h = 31.1 min *[8]*). Thus, in addition to genes "essential for survival" (that cause growth arrest or cell death upon inactivation), a fraction of genes required to maintain a robust growth at the rate of 40% (or better) than that of the wild type ("essential for fitness") were asserted as "essential" in this experiment. Detailed comparison of this experimental data set with the list of essential and dispensable genes determined by systematic gene inactivation approach with clonal outgrowth *(2)* indirectly confirms this conclusion (*[4]* and **Note 4**).

To test the influence of a readout sensitivity threshold on assessment of gene essentiality, let us consider a hypothetical experiment in which the detection limit is an order of magnitude higher than that in **equation 10**:

$$\beta_{x2} = \frac{1}{12}\left(\ln\left(1.4\times10^{9}\right) + \ln\left(10^{-4}/22\right)\right) = 0.730.$$

This corresponds with the doubling time of 0.95 h (**equation 2**). Hence, an essentiality study with the populational outgrowth and a poor readout sensitivity will result in scoring as "lethal" all mutants capable of growth rates up to ~55% of that of the wild type, leading to an overestimation of the number of essential genes.

To estimate the influence of the outgrowth time on the assessment of gene essentiality, let us consider a hypothetical experiment in which the outgrowth period was shorter (9 h, ~17 population doublings) than that described in **equation 10**:

$$\beta_{x3} = \frac{1}{9}\left(\ln\left(4.8 \times 10^{6}\right) + \ln\left(10^{-5}/22\right)\right) = 0.087$$

This corresponds with the doubling time of 8 h (6.5% of the wild type growth rate, nearly complete growth arrest). Hence, at significantly shorter propagation times, only genes "essential for survival" will be detected. Note, however, that **approximation 8**, although useful, does not account for other factors potentially important in planning a genome-scale gene essentiality experiment, such as probable persistence in the population of genomic DNA from nonviable cells (if short outgrowth times are used). Highly sensitive detection techniques can erroneously register such DNA fragments as viable mutants, leading to false negatives in scoring essential genes. It is recommended that the optimal number of population doublings necessary to reduce the titer of cells with insertions in core essential genes beyond detection level be determined in a pilot study, using **approximation 8** as a starting point.

Notes

1. This simplification is justified in our example because cells were propagated only to a late logarithmic stage.
2. The sensitivity threshold is generally determined in each experiment by preparing a battery of serial dilutions of a test mutant within a population and processing these control samples by the standard detection procedure used in the large-scale experiment.
3. The total number of 2×10^5 independent insertion mutants generated in this experiment corresponds with 1 insert per 46 bp of genomic sequence, which amounts to approximately 22 inserts per an average (1000 bp long) *E. coli* ORF ***(8)***.
4. In *E. coli* where both populational ***(8)*** and clonal ***(2)*** essentiality screens were conducted on a genome scale, the two corresponding data sets reveal discrepancies in the essentiality assignments of 437 genes (~12% of the 3580 genes unambiguously assigned in both projects). Among them, 393 genes (~10.8% of the common set, or 90% of the differing 437 assignments) were deemed essential by populational screen but dispensable in the clonal collection. Notably, the discrepancies of the opposite kind were observed only for 44 genes (1.2% of the set). Thus, the populational essentiality screen has successfully identified nearly all the genes "essential for survival" but has also scored 393 additional genes strongly contributing to robust cellular growth.

References

1. Winterberg, K. M., Luecke, J., Bruegl, A. S., and Reznikoff, W. S. (2005) Phenotypic screening of *Escherichia coli* K-12 Tn5 insertion libraries, using whole-genome oligonucleotide microarrays. *Appl. Environ. Microbiol.* **71**, 451–459.
2. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., et al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knock-out mutants: the Keio collection. *Mol. Systems Biol.*

3. Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M. R., Maruf, M., et al. (2006) Essential genes of a minimal bacterium. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 425–430.
4. Gerdes, S., Edwards, R., Kubal, M., Fonstein, M., Stevens, R., and Osterman, A. (2006) Essential genes on metabolic maps. *Curr. Opin. Biotechnol.* **17**, 448–456.
5. Smith, H. L., and Waltman, P. (1995) *The Theory of the Chemostat.* Cambridge: Cambridge University Press.
6. Passarge, J., and Huisman, J. (2002) Competition in well-mixed habitats: From competitive exclusion to competitive chaos. In: Sommer, U. and Worm, B., eds. *Competition and Coexistence. Ecological Studies*, vol. 161. New York: Springer, pp. 7–42.
7. Hesseler, J., Schmidt, J. K., Reichl, U., and Flockerzi, D. (2006) Coexistence in the chemostat as a result of metabolic by-products. *J. Math. Biol.* **53**, 556–584.
8. Gerdes, S., Scholle, M., Campbell, J., Balazsi, G., Ravasz, E., Daugherty, M., et al. (2003) Experimental determination and system-level analysis of essential genes in *E. coli* MG1655. *J. Bacteriol.* **185**, 5673–5684.
9. Gompertz, B. (1825) On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philos. Trans. R. Soc. London* **115**, 513–585.
10. Zwietering, M. H., Jongenburger, I., Rombouts, F. M., and Vantriet, K. (1990) Modeling of the bacterial-growth curve. *Appl. Environ. Microbiol.* **56**, 1875–1881.

25

Statistical Analysis of Fitness Data Determined by TAG Hybridization on Microarrays

Brian D. Peyser, Rafael Irizarry, and Forrest A. Spencer

Summary

TAG, or bar-code, microarrays allow measurement of the oligonucleotide sequences (TAGs) that mark each strain of deletion mutants in the *Saccharomyces cerevisiae* yeast knockout (YKO) collection. Comparison of genomic DNA from pooled YKO samples allows estimation of relative abundance of TAGs marking each deletion strain. Features of TAG hybridizations create unique challenges for analysis. Analysis is complicated by the presence of two TAGs in most YKO strains and the hybridization behavior of TAGs that may differ in sequence from array probes. The oligonucleotide size of labeled TAGs also results in difficulty with contaminating sequences that cause reduced specificity. We present methods for analysis that approach these unique features of TAG hybridizations.

Key Words: bar code, deletion, knockout, microarray, *Saccharomyces cerevisiae*, TAG, yeast.

1. Introduction

TAGs are unique oligonucleotides that serve as molecular identifiers to detect the presence of a specific DNA molecule. During the creation of the knockout collection for the budding yeast *Saccharomyces cerevisiae*, two TAGs were incorporated into the design of each null allele (*[1, 2]* and **Chapter 15**). These flank the drug-resistance cassette that replaces an open reading frame. Each yeast knockout (YKO) allele was assigned a unique UPTAG and DNTAG residing "upstream" and "downstream" of the drug-resistance cassette. The TAGs themselves are placed between primer binding sites suitable for polymerase chain reaction (PCR) amplification. Thus, populations of TAGs may be amplified using universal primers for all UPTAGs, or distinct universal primers for all DNTAGs, in a DNA sample.

TAG microarrays allow indirect measurement of YKO strain representation for all mutants simultaneously *(1, 2)*. In brief, genomic DNAs from experimental and control cultures are used to template a pair of PCR reactions that generate labeled TAG oligonucleotides. Four reactions form a complete set (UPTAG control, UPTAG experiment, DNTAG control, and DNTAG experiment). The four products may be mixed for

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

cohybridization in a two-color experiment (e.g., Agilent [3]), or experiment and control samples may be hybridized on two separate arrays in a single label protocol (e.g., Affymetrix [4, 5]).

Microarray-based detection of individual elements within a complex pool has been employed in budding yeast as a means to measure the relative fitness of cells bearing different knockout mutations (reviewed in Ref. 6). The approach was pioneered during development of the yeast deletion collection (1, 7) and has been used to characterize growth defects in response to a battery of culture conditions (2) and for drug sensitivity profiling (8–13). TAG arrays have also been used to identify haploinsufficient loci (14) as well as interacting pairs of mutants that show synthetic fitness or lethal phenotypes ([3, 5, 15–17] and **Chapter 15**). The TAG microarray approach is appropriate for many experimental designs where the relative representation of YKOs in control and experimental populations are to be compared.

TAG arrays present unique problems for researchers due to mismatches between TAG sequences contained in the actual strain and the "correct" TAG sequences on the microarray. Although many of the TAG sequence discrepancies for the heterozygous deletion collection are known, this knowledge is insufficient to fully explain hybridization behavior (18, 19). Additionally, the presence of two TAGs in almost every YKO strain results in challenges and opportunities for analysis. Here we describe statistical methods for detecting the relative representation of YKO strains under experimental and control conditions.

2. Materials

1. TAG microarray results file and documentation.
2. Spreadsheet software.
3. Web browser and internet connection.

3. Methods

3.1. Log-Transform

Whenever visualizing results from a microarray, log-transformed data typically provide more informative plots than untransformed data (**Fig. 1**). We suggest log base 2 because it is easy to mentally translate. For example, 5 is $\log_2 32$, and 10 is $\log_2 1024$. Others have introduced the "glog" ("generalized logarithm"), which is motivated by a transformation that removes the dependence of relative intensity variance on the mean intensity (20–22). There are several advantages to this transformation; however, it is related to the *arcsinh* function and has the disadvantage that it is not easy to mentally translate values to the original scale. Additionally, the glog requires more sophisticated analysis tools.

A useful visualization using log intensities is the M versus A or ratio-intensity plot (23). This plot shows the log(ratio) versus the average log(intensity) (**Fig. 2**). Because $\log_2 R/G$ is equal to $\log_2 R$ minus $\log_2 G$, the log ratio is easily calculated as the difference between log intensities. **Figure 2** shows the same data as **Figure 1**; however, the slight dependence of log ratios on intensity is more evident. Notice in **Figure 2** that the high-intensity data exhibit a general tendency toward negative log ratio values.

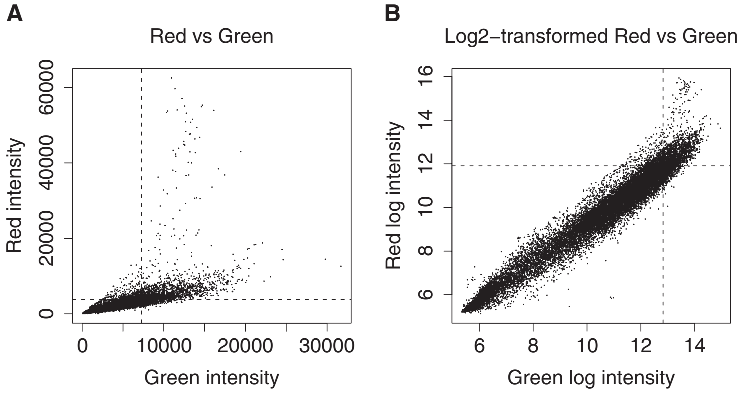


Fig. 1. Comparison of signal intensity plots using raw and log2 transformed data. **(A)** Red (Cy5) intensity versus green (Cy3) intensity for a two-color array. **(B)** Red $\log_2$(intensity) versus green $\log_2$(intensity) for a two-color array. Dashed lines are 90th percentile for each color; therefore, 81% of all data are in the lower left quadrants.

3.2. Background Correction

There is evidence that local background intensities affect measurements of spot intensity. Manufacturers typically provide image-analysis software that supplies estimates of these local background levels and a default correction based on the difference between the observed intensity and observed background level. However, we find these measurements imperfect at best. Whereas the sensitivity of TAG arrays can be increased by subtracting background values, often, specificity is too heavily impacted (**Fig. 3A**). Statistical models predict this increase in variance ***(22, 24)***. In particular, this variance-explosion effect occurs at low intensities even with synthetic data plus appropriate noise. Additionally, some features report background levels higher than foreground levels, resulting in negative background–corrected values. Because the logarithm is

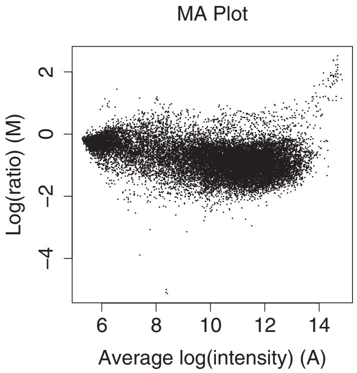


Fig. 2. MA or ratio-intensity plot. The difference between log intensities: $\log_2 R - \log_2 G = \log_2 R/G$ versus the average log intensity: $(\log_2 R + \log_2 G)/2 = \log_2 \sqrt{(R \times G)}$. The data are from **Fig. 1**.

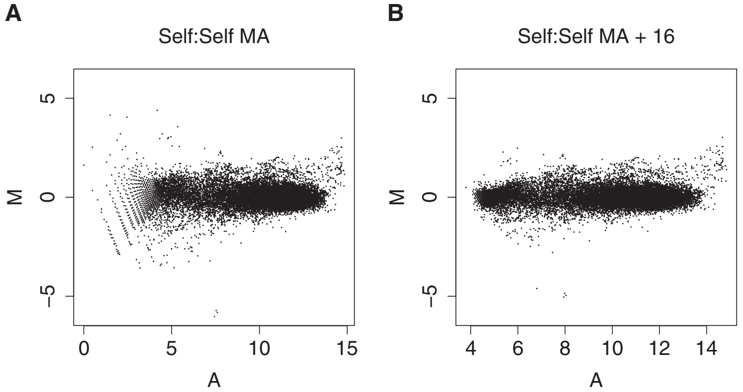


Fig. 3. Background value subtraction causes increased variance at low intensities. **(A)** MA plot of self versus self hybridization with intensities corrected by subtracting the value provided by image-processing software (GenePix; Molecular Devices, Sunnyvale, CA). Twenty-nine points are missing from this plot because background-corrected values are zero or negative, resulting in undefined logarithms. **(B)** MA plot after addition of 16 ($\log_2 16 = 4$) to all background-corrected values. No points are missing.

defined only for positive numbers, this results in loss of data. One important distinction that may explain the failure of background measurements is that they are usually taken from images of glass surface with no oligonucleotide or cDNA attached. The actual background levels for features may differ from nearby array surfaces due to the presence of DNA at the feature.

Improvement on the imperfect measurements provided by standard image-analysis software is provided by model-based background–correction methods such as limma (*25*), vsn (*26*), RMA (*27*), or CRAM (Yuan and Irizarry). However, these procedures require familiarity with statistical open source software such as R (http://www.r-project.org/, see also **Chapter 22**) and BioConductor (http://www.bioconductor.org/). We highly recommend learning this software as it provides the most powerful analysis tools available. However, for those who would rather point and click, commercial packages exist that implement some of these procedures. Even with these, it will be difficult to work with the UPTAG and DNTAG separately, which we will show is necessary (**Section 3.3**). Therefore, we suggest a simple ad hoc procedure for overcoming these problems with background correction and log-transformation. Addition of a fixed value to all background-subtracted features can solve the variance-explosion and undefined logarithm problems. We show addition of 16 to all background-adjusted values in **Figure 3B**.

3.3. Microarray Data Normalization

3.3.1. Common Strategies for Data Normalization

It is common to see intensity-dependent biases in the observed log-ratios (**Fig. 2**). Various normalization methods exist to correct these biases. They depend on the assumption that either an approximately equal number of strains decrease and increase in representation between conditions or that most strains do not change. If these assumptions are not plausible for your experiment, we recommend that you consult an expert in analysis of microarray data to identify an appropriate approach rather than relying on widely available methods.

Spreadsheet applications are not capable of statistics required for two-color normalization *(28)*. An easy-to-use tool for performing normalization is SNOMAD (standardization and normalization of microarray data), available on the Web *(29)*. We recommend this normalization procedure (called "Local Mean Normalization Across Element Signal Intensity") for two-color microarrays (**Note 1**). The procedure involves passing a line through a plot of the log ratio (*M*) versus average log signal intensity (*A*) *(30)*. The line is produced using a robust fit procedure called "loess" *(31)*. The log ratios are then corrected to the residuals so that the line is horizontal at $M = 0$.

The loess normalization procedure is available in many commercial microarray-analysis packages. However, UPTAG and DNTAG hybridizations result from independent labeling reactions and therefore must be normalized separately (**Fig. 4**). A "normalize checkbox" in commercial packages will likely inappropriately normalize UPTAG and DNTAG simultaneously. The lines in **Figure 4** are calculated with "Span" = 0.3; otherwise, the fits mirror the default settings of SNOMAD. "Span" values closer to 1 result in a smoother fit, as it is the fraction of data used to estimate the local fit along *A* values. If microarrays are spotted by print-tips, it may also be useful to subdivide the array by print-tip in addition to subdividing by TAG type *(30)*. This procedure (print-tip loess) is implemented in some commercial packages and can be performed using SNOMAD by passing each print-tip group of UPTAGs and DNTAGs separately.

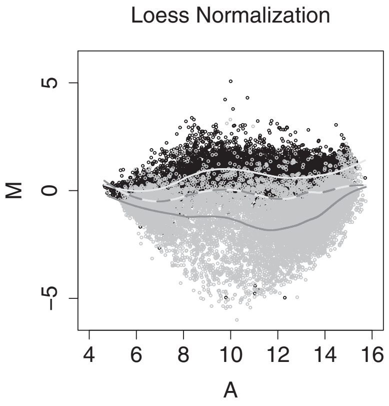


Fig. 4. MA plot. Log ratios are plotted by average log intensity, and a loess curve is shown for each subarray. DNTAG values are plotted in black, and the corresponding loess curve is light gray. UPTAG values are gray, with corresponding loess curve in dark gray. The incorrect loess curve produced using all data simultaneously instead of separately is a light/dark gray dashed line. If the data were incorrectly normalized simultaneously, most DNTAGs would have positive log ratios, whereas most UPTAGs would have negative log ratios in this example.

3.3.2. Using SNOMAD for Data Normalization

1. In order to use SNOMAD for normalization, first perform background correction as described above by subtracting the manufacturer-supplied estimate and then adding 16. We recommend using the median pixel intensity for both foreground and background, as it is less susceptible to noise/artifacts (such as dust) than the mean intensity.
2. Once the background-corrected red and green values are obtained, separate the results into DNTAG and UPTAG arrays. At this point, it is important to be sure that later realignment of UPTAG and DNTAG data from the same strain is easy. Use of open reading frame (ORF) annotation (e.g., YAL001C) is not sufficient because duplicate knockouts exist for a number of ORFs. Additionally, a small number of YKOs lack DNTAGs, so for these only the UPTAG information is available. The best approach is to annotate each feature with a strain-specific identifier so that corresponding UP- and DNTAGs can be properly combined. Split the data into two spreadsheets, one for UPTAG values and one for DNTAG values. The background-adjusted red and green intensities might be contained in columns 1 and 2 for this example. There may be an unlimited number of additional columns for annotation if desired, but only the two background-adjusted intensity columns will be used by the SNOMAD software.
3. Save each UP- or DNTAG sheet in tab-delimited text format and upload the data to SNOMAD (http://pevsnerlab.kennedykrieger.org/snomadinput.html). Choose the UP- or DNTAG file to upload on the right side of the Web page and specify the columns for background-adjusted red and green values (ONEintensities and TWOintensities), that is, columns 1 and 2.
4. Scroll down to section three, "Logarithmic Transformation," check "Perform This Transformation," and input *log base 2*.
5. In section four, choose to perform "Calculate Mean Log(Intensities) and Log(Ratios)." This will create an MA plot, which you can view if desired.
6. Last, perform "Local Mean Normalization Across Element Signal Intensity" with "Span" set to 0.3 and default "Trim" of 0.1. At the bottom of the page, optionally choose a file name and click "Submit Data for Processing."
7. The tab-delimited text file returned by SNOMAD will contain six new columns at the end of your submitted data. The "meanlogint" column contains the average log intensities (A), and the "logratiores" column contains the corrected (normalized) log ratios (M).
8. To convert the M and A values to corrected red and green log intensities, add (red) or subtract (green) one-half of the M value from the A value (because $M = \log_2 R - \log_2 G$ and $A = (\log_2 R + \log_2 G)/2$; $M = 2 \times A - 2 \times \log_2 G$, therefore, $\log_2 G = A - M/2$).

3.4. Identifying Nonfunctional TAGs

A straightforward, but naïve, approach to summarizing the UP- and DNTAG measurements is to simply average the observed log ratios. However, this method will reduce sensitivity when one of the TAGs is nonfunctional. Many TAG sequence discrepancies have been identified *(18)*, where the oligonucleotide incorporated into the YKO strain is not identical to the one intended and therefore differs from the one that has been designed as an array feature. Presumably, these observations reflect synthesis errors fixed by cloning of an individual molecule at transformation during generation of a given YKO strain. However, TAG hybridization behavior is not fully predicted by knowledge of the presence and nature of mutations *(18, 19)*. TAG hybridization failure is readily apparent in histograms of corrected $\log_2$ signal intensity from a pooled heterozygous deletion collection. Although all YKO alleles are present, TAG signals display a bimodal distribution (**Fig. 5**). The lower peak is close to background of $\log_2 16 = 4$ (**Section 3.2**) and consists of TAGs that hybridize poorly.

Steps taken in analysis can minimize the impact of poorly hybridizing TAGs. Non-hybridizing TAG behavior can be defined using the array data to define threshold values for TAG intensities representing "present" or "absent" hybridization signal. Known negative control features are very useful in providing an empirical measurement of signal intensities that correspond with "absent" hybridization. On the "Hopkins TAG Array" (GEO accession GPL1444), the "YQL" features ***(19)*** can be used as negative controls. On other TAG arrays, researchers could use features representing essential YKOs if the sample is from a haploid or homozygous deletion pool.

A threshold value for each array can be calculated using the median plus three standard deviations (SD) for $\log_2$ Cy5 intensity values of negative control features (**Note 2**). The Cy5 background-corrected and normalized intensities can be calculated as: $M + 0.5 \times A$. Because the negative control intensity data are highly skewed, use a robust estimate of the SD: the median absolute deviation (MAD). This value is simple to calculate even with spreadsheet software. For all negative control $\log_2$ intensity values x, calculate the absolute value of x minus the median negative control $\log_2$ intensity. These are the "absolute deviations." Then find the median of these values and multiply by 1.4826 (approximately $\varphi^{3/4}$, or the inverse of the value at which the cumulative distribution $\phi(x) = 0.75$), so that the MAD is on the same scale as the SD. Thus, if negative control $\log_2$ Cy5 intensity measurements are in spreadsheet cells A1 through A800, the Excel syntax for calculation of MAD is

```
  B1 = ABS(A1-MEDIAN(A$1:A$800))   (repeat for A2:A800 in cells B2:B800)
MAD = 1.4826*MEDIAN(B1:B800).
```

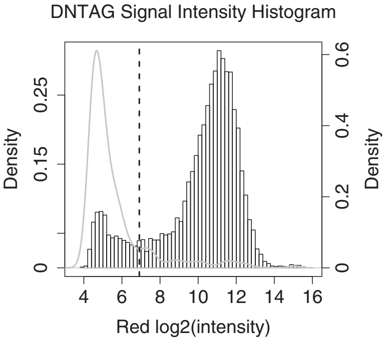


Fig. 5. DNTAG Cy5 intensity distribution. Histogram of corrected $\log_2$ DNTAG Cy5 intensity for features that correspond with TAGs. UPTAG and Cy3 are similar. All features should be present; however, some features hybridize near the same intensity as negative control spots. Gray line and right axis show intensity distribution of negative control features, which are similar to TAGs but not present in any YKO (*see* Ref. ***19***). Dashed line is at median plus 3 MAD (robust SD) for negative control features. TAGs with Cy5 $\log_2$ intensity values smaller than shown by the dashed line are annotated "absent" and removed when the corresponding UPTAG is present.

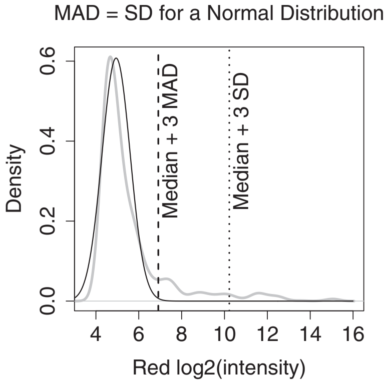


Fig. 6. MAD = SD for normal distributions. The gray line depicts the distribution of negative control features from **Figure 5**, and the black line shows the normal distribution where the mean/median and SD are equal to the median and MAD of the negative control distribution. Because the negative control distribution has a large right tail, the SD is larger than the MAD (dotted line vs. dashed line). For the normal distribution shown in black, the SD and MAD are equal, and both measurements yield the dashed line.

The MAD scaled this way is identical to the SD for normally distributed data. Data from negative control spots on TAG arrays are typically skewed with a long right tail and result in a larger SD than MAD (**Fig. 6**).

Once the threshold value for a given array is determined, each UPTAG and DNTAG can be annotated "present" (probability UP/DN is present = $p(UP/DN\ present) = 1$) or "absent" ($p(UP/DN\ present) = 0$) based on the TAG intensity. The corresponding log ratios are averaged for each YKO using $w \times UP + (1 - w) \times DN$, where $w = 0.5 + (p(UP\ present) - p(DN\ present))/2$. This results in a weight w of 0, 0.5, or 1 when UP or DN can be only "present" (1) or "absent" (0). Thus, the weighted average log ratio is the average of UP and DN if both are present or both are absent, or it is only UP if it is present while DN is absent (and vice versa). For YKOs with no DNTAG, always use UPTAG. **Note 3** describes alternative approaches to identification and discounting of nonhybridizing TAGs.

3.5. Artifact Filtering

Artifact signal on TAG arrays is derived from several sources. TAGs present as contaminants in the labeling PCR can impact both sensitivity and specificity. To minimize contamination, it is important to prepare all reagents that enter the labeling reaction carefully to avoid unwanted introduction of template from ambient lab sources (*see* Ref. ***19*** and **Chapter 15**). However, some fluorescent signal can be observed on microarray features even in the absence of a labeling reaction (**Fig. 7**), indicating that it is intrinsic to a reagent. Additionally, some TAGs hybridize more strongly in a single channel of two-channel data routinely (i.e., across many independent array experiments). This is consistent with contamination of one primer set. Artifacts like this have been found reproducibly within a single primer set batch, but different TAGs misbehave with new batches of labeled primers ***(32)***.

An efficient method (**Note 4**) for recognizing primer batch–specific artifacts is a self: self hybridization, where the same DNA is used as template for both dye sets. This produces a known ratio of 1 for all TAGs, and any data that deviate from expected can be recognized and filtered from all sets. After normalization, in self:self experiments most YKO strains show the expected bivariate normal distribution about log ratio of zero ($\log_2 1 = 0$) for UPTAG and DNTAG, but some strains differ substantially. A useful filter removes outliers, defined as TAGs with log ratios greater than three MAD from zero (*see* example in **Fig. 8**). After generating the weighted average of UP and DNTAG log ratios, use the artifact filter to remove data that are outliers in self:self hybridizations. For example: if a DNTAG was annotated "absent," but the UPTAG deviates from zero in self:self hybridizations, replace the log ratio with the DNTAG value. This favors reduced sensitivity of a poorly hybridizing DNTAG over the low specificity of an artifact UPTAG. If both UP- and DNTAG display artifact signal, remove the strain from consideration by filtering both TAGs. Similarly remove from consideration any YKOs with only UPTAG when it exhibits artifact hybridization.

3.6. Ratio Determination and Data Interpretation

The data preprocessing **Sections 3.1** to **3.5** are recommended for generating good measurements of YKO strain representation in a population using TAG microarray

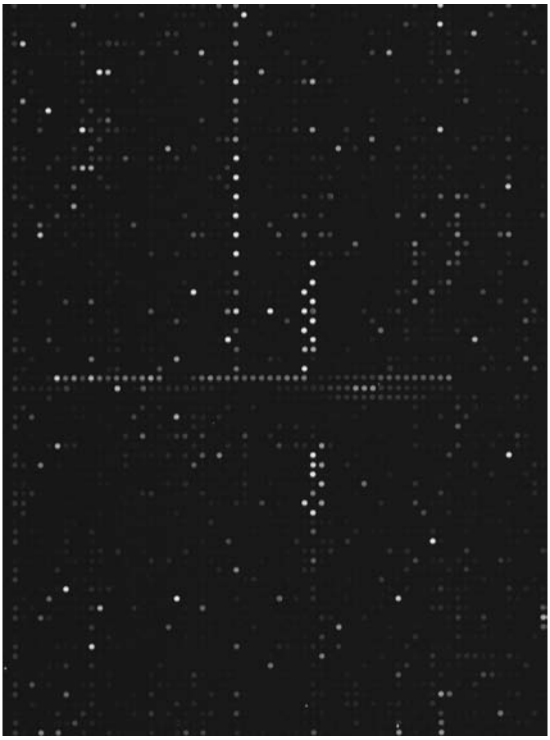


Fig. 7. Hybridization of UPTAG and DNTAG primers. A section of a microarray hybridized with labeled primers not subjected to PCR.

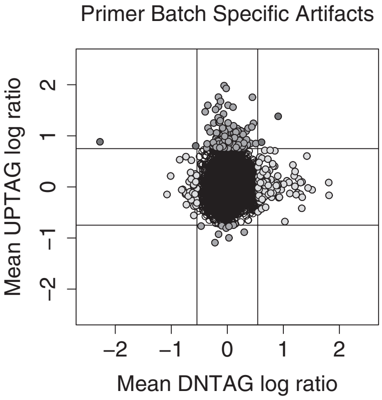


Fig. 8. Self:self arrays. Mean UP- and DNTAG log ratios from seven self:self hybridizations performed using a single set of primers. Lines are 3 MAD from zero. Light gray points, YKOs with DNTAG log ratio > 3 MAD from zero; gray points, YKOs with UPTAG log ratio > 3 MAD from zero; dark gray points, YKOs with both UPTAG and DNTAG > 3 MAD from zero.

data. Ensuing analyses of those values from control and experimental data sets will be driven by specific experimental designs, a typical one being ranking by experimental: control ratio to reveal population representation difference for each YKO under two conditions being compared.

As noted, TAG array data reflect several known sources of noise, some of which can be minimized using appropriate analysis strategies. A spike-in experiment that varied the representation of yeast TAGs at known ratios indicates that published hybridization and analysis procedures can yield quite good detection *(31)*. In practice, the presence of false-positive observations from noise limits the sensitivity on true-positive identification (*see*, e.g., a true-positive distribution in Ref. 5). The effect of biological sample handling, such as control versus experimental selection efficiencies, may continue to be a major source of false positives and is worthy of careful attention.

Future improvement in the analysis of TAG arrays may come from better understanding of TAG hybridization behavior through large-scale multiarray analyses (R.A. Irizarry, B.D. Peyser, and D.S. Yuan, unpublished). The procedures provided here are intended to be used with small or large numbers of array data sets. They are applicable to any experimental design where YKO representation is measured as a function of UPTAG and DNTAG signal intensity.

Notes

1. For one-color analyses, a normalization procedure called “quantile normalization” is a good choice *(33)*. The normalization procedure is computationally simple and can even be performed manually with a spreadsheet application. This procedure usually does not perform as well on two-color arrays as a loess procedure because array-specific location effects (e.g., scratches) are decoupled, but it is useful for comparisons across arrays. To perform

quantile normalization manually with spreadsheet software, first separate the UPTAG and DNTAG subarrays. Being sure to keep feature identification with signal intensities, sort the intensities of each array from lowest to highest. Then average the intensity across all UP- or DNTAG subarrays for the first (smallest) value, second value, and so forth, and replace the feature value on each array with the average across arrays. The normalized values are then log-transformed and reordered by feature ID so that comparisons across arrays can be made. This procedure results in each array sharing exactly the same distribution of values, but the identities of features corresponding with each value differ.

2. The experimental design will determine which channel(s) are best to use for finding nonfunctional TAGs. It is useful to bear in mind that defining a threshold will not always perfectly separate "absent" from "present" and can cause a subset of true low-hybridization signals to be discounted as "absent." The extent of this effect will depend on overlap between present and absent signal intensity distributions. For example, a Cy5-based negative control procedure will introduce a bias against recognizing some TAGs that exhibit decreased representation in Cy5- relative to Cy3-labeled extract. This will cause some number of false-positive or false-negative conclusions to be drawn, depending on the role of the Cy5-labeled extract in the experimental design (as control or experimental sample). Thus, the decision to use Cy5, Cy3, or their average for defining nonfunctional TAGs can be based on whether the potential for false-positive or false-negative observations is best tolerated. An additional source of artifact may come from the presence of cross-hybridization by a small subset of TAGs. For example, if a DNTAG cross-hybridizes to give high signal in both experimental and control samples, and the corresponding UPTAG appropriately represents abundance but in a low-signal-intensity range, the use of an "absence" threshold will augment use of data from the cross-hybridizing TAG. It has been estimated that 3% to 5% of TAGs exhibit significant cross-hybridization ***(19)***.
3. Several methods for finding nonfunctional TAGs are possible. The method presented here is straightforward and simple to implement, uses data internal to the specific experiment, and is appropriate for any TAG array hybridization regardless of the YKO source. A variation on this method produces a continuous scale weight between 0 and 1 for each UP/DNTAG based on the probability each TAG is present ***(31)***. A simpler approach is to use a previously derived list of nonfunctional TAGs to filter data from all experiments. For researchers using the heterozygous YKO collection, a list of nonfunctional TAGs is available from The Johns Hopkins University Genetic Interaction Map of Yeast project (http://slam.bs.jhmi.edu/). This list is based on behavior of TAGs across many (in the range of 10^3) hybridizations. An advantage of using this list is that it does not bias toward an UP/DNTAG that cross-hybridizes when the corresponding DN/UPTAG accurately displays low signal for a non-abundant strain (**Note 2**). Note that the actual TAGs present (and identity of nonfunctional TAGs) in the various haploid and diploid YKO collections will vary due to the fact that multiple independent transformation events lead to the establishment of a given YKO allele across these collections.
4. An alternative method for approaching TAG-specific hybridization artifacts is to perform a "dye-swap" experiment for each pair of samples, where the Cy5 and Cy3 primers are used to label the opposite DNA samples. The UP- and DNTAG log ratios may then be averaged across the two dye orientations to remove primer artifacts. The disadvantage of this approach is that it doubles the number of hybridizations required. When a single experimental condition is interrogated, this is not a large expense, but when many hybridizations are planned, it may be more efficient to perform control self:self hybridizations to characterize outliers specific to a primer batch.

Acknowledgments

The authors thank D.S. Yuan, J.D. Boeke and P.B. Meluh for discussion, O. Chen and C. Tiffany for expert technical support, and X. Zhao for comments on the manuscript. This work was supported by NIH grants GM62368, HG02432, and GM007445.

References

1. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M., and Davis, R. W. (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat. Genet.* **14**, 450–456.
2. Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391.
3. Pan, X., Yuan, D. S., Xiang, D., Wang, X., Sookhai-Mahadeo, S., Bader, J. S., et al. (2004) A robust toolkit for functional profiling of the yeast genome. *Mol. Cell* **16**, 487–496.
4. Winzeler, E. A., Castillo-Davis, C. I., Oshiro, G., Liang, D., Richards, D. R., Zhou, Y., and Hartl, D. L. (2003) Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics* **163**, 79–89.
5. Warren, C. D., Eckley, D. M., Lee, M. S., Hanna, J. S., Hughes, A., Peyser, B., et al. (2004) S-phase checkpoint genes safeguard high-fidelity sister chromatid cohesion. *Mol. Biol. Cell* **15**, 1724–1735.
6. Ooi, S. L., Pan, X., Peyser, B. D., Ye, P., Meluh, P. B., Yuan, D. S., et al. (2006) Global synthetic-lethality analysis and yeast functional profiling. *Trends Genet.* **22**, 56–63.
7. Winzeler, E. A., Lee, B., McCusker, J. H., and Davis, R. W. (1999) Whole genome genetic-typing in yeast using high-density oligonucleotide arrays. *Parasitology* **118** (Suppl), S73–80.
8. Giaever, G., Shoemaker, D. D., Jones, T. W., Liang, H., Winzeler, E. A., Astromoff, A., and Davis, R. W. (1999) Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nat. Genet.* **21**, 278–283.
9. Giaever, G. (2003) A chemical genomics approach to understanding drug action. *Trends Pharmacol. Sci.* **24**, 444–446.
10. Giaever, G., Flaherty, P., Kumm, J., Proctor, M., Nislow, C., Jaramillo, D. F., et al. (2004) Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 793–798.
11. Lum, P. Y., Armour, C. D., Stepaniants, S. B., Cavet, G., Wolf, M. K., Butler, J. S., et al. (2004) Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. *Cell* **116**, 121–137.
12. Dunn, C. D., Lee, M. S., Spencer, F. A., and Jensen, R. E. (2006) A genomewide screen for petite-negative yeast strains yields a new subunit of the i-AAA protease complex. *Mol. Biol. Cell* **17**, 213–226.
13. Arevalo-Rodriguez, M., Pan, X., Boeke, J. D., and Heitman, J. (2004) FKBP12 controls aspartate pathway flux in *Saccharomyces cerevisiae* to prevent toxic intermediate accumulation. *Eukaryot. Cell* **3**, 1287–1296.
14. Deutschbauer, A. M., Jaramillo, D. F., Proctor, M., Kumm, J., Hillenmeyer, M. E., Davis, R. W., et al. (2005) Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* **169**, 1915–1925.
15. Ooi, S. L., Shoemaker, D. D., and Boeke, J. D. (2003) DNA helicase interaction network defined using synthetic lethality analyzed by microarray. *Nat. Genet.* **35**, 277–286.

16. Lee, M. S., and Spencer, F. A. (2004) Bipolar orientation of chromosomes in *Saccharomyces cerevisiae* is monitored by Mad1 and Mad2, but not by Mad3. *Proc. Natl Acad. Sci. U.S.A.* **101**, 10655–10660.
17. Pan, X., Ye, P., Yuan, D. S., Wang, X., Bader, J. S., and Boeke, J. D. (2006) A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell* **124**, 1069–1081.
18. Eason, R. G., Pourmand, N., Tongprasit, W., Herman, Z. S., Anthony, K., Jejelowo, O., et al. (2004) Characterization of synthetic DNA bar codes in *Saccharomyces cerevisiae* gene-deletion strains. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11046–11051.
19. Yuan, D. S., Pan, X., Ooi, S. L., Peyser, B. D., Spencer, F. A., Irizarry, R. A., and Boeke, J. D. (2005) Improved microarray methods for profiling the Yeast Knockout strain collection. *Nucleic Acids Res.* **33**, e103.
20. Huber, W., von Heydebreck, A., Sueltmann, H., Poustka, A., and Vingron, M. (2003) Parameter estimation for the calibration and variance stabilization of microarray data. *Stat. Appl. Genet. Mol. Biol.* **2**, article 3.
21. Durbin, B. P., Hardin, J. S., Hawkins, D. M., and Rocke, D. M. (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* **18** (Suppl 1), S105–110.
22. Durbin, B. P., and Rocke, D. M. (2004) Variance-stabilizing transformations for two-color microarrays. *Bioinformatics* **20**, 660–667.
23. Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–139.
24. Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004) A model-based background adjustment for oligonucleotide expression Arrays. *J. Am. Statist. Assoc.* **99**, 909–917.
25. Smyth, G. K. (2005) Limma: linear models for microarray data. In: Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., and Huber, W., eds. *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. New York: Springer, pp. 390–420.
26. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002) Variance stablization applied to microarray data calibration and to quantification of differential expression. *Bioinformatics* **18** (Suppl 1), S96–104.
27. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.
28. Yuan, D. S., and Irizarry, R. A. (2006) High-resolution spatial normalization for microarrays containing embedded technical replicates. *Bioinformatics* **22**, 3054–3060.
29. Colantuoni, C., Henry, G., Zeger, S., and Pevsner, J. (2002) SNOMAD (Standardization and Normalization of MicroArray Data): WEB-accessible gene expression data analysis. *Bioinformatics* **18**, 1540–1541.
30. Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15.
31. Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Statist. Assoc.* **74**, 829–836.
32. Peyser, B. D., Irizarry, R. A., Tiffany, C. W., Chen, O., Yuan, D. S., Boeke, J. D., and Spencer, F. A. (2005) Improved statistical analysis of budding yeast TAG microarrays revealed by defined spike-in pools. *Nucleic Acids Res.* **33**, e140.
33. Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193.

26

Profiling of *Escherichia coli* Chromosome Database

Yukiko Yamazaki, Hironori Niki, and Jun-ichi Kato

Summary

The Profiling of *Escherichia coli* Chromosome (PEC) database (http://www.shigen.nig.ac.jp/ecoli/pec/) is designed to allow *E. coli* researchers to efficiently access information from functional genomics studies. The database contains two principal types of data: gene essentiality and a large collection of *E. coli* genetic research resources. The essentiality data are based on data compilation from published single-gene essentiality studies and on cell growth studies of large-deletion mutants. Using the circular and linear viewers for both whole genomes and the minimal genome, users can not only gain an overview of the genome structure but also retrieve information on contigs, gene products, mutants, deletions, and so forth. In particular, genome-wide exhaustive mutants are an essential resource for studying *E. coli* gene functions. Although the genomic database was constructed independently from the genetic resources database, users may seamlessly access both types of data. In addition to these data, the PEC database also provides a summary of homologous genes of other bacterial genomes and of protein structure information, with a comprehensive interface. The PEC is thus a convenient and useful platform for contemporary *E. coli* researchers.

Key Words: essential genes; minimum genome; mutant strains.

1. The Profiling of *Escherichia coli* Chromosome Database Structure and Content

The Profiling of *Escherichia coli* Chromosome (PEC) database contains substantial information for *E. coli* researchers. It includes (1) gene essentiality data, (2) minimal genome and large-deletions information, (3) structural features (domain, motif, etc.) of each gene product, (4) results of comparative analysis of *E. coli* genes and their homologues in other bacterial genomes, and (5) strain and plasmid collections. PEC provides (6) a circular and linear genome viewer, (7) BLAST service and tools, (8) various downloadable files, and (9) PEC contigs of *E. coli* MG1655 genomic sequence, U00096.2. The logical schema of the PEC database is shown in **Figure 1**.

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

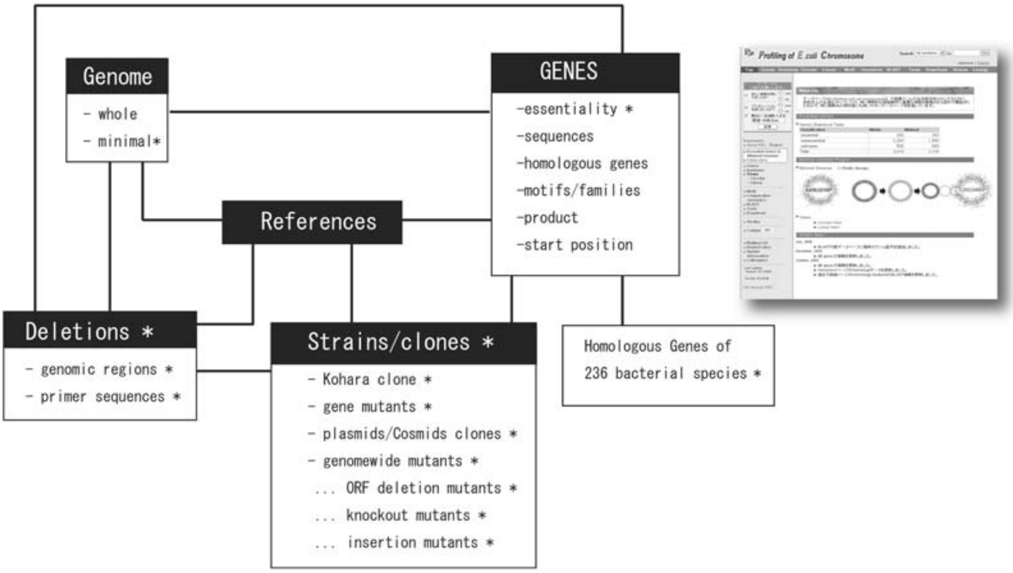


Fig. 1. Logical schema (left) and the front page (right) of the PEC database. Asterisks indicate the original PEC contents.

1.1. Gene Essentiality Data in PEC

All *E. coli* genes were classified into three groups based on the information obtained from the literature: (1) essential for cell growth, (2) dispensable for cell growth, and (3) unknown. Gene classification was based on the following criteria.

1.1.1. Experimental Evidence Is Available

Basically, if a strain harboring a null-type mutation in a gene (in the absence of suppressor mutations) was able to grow, the gene was classified as "nonessential", even if the strain could grow only at a certain temperature or under certain nutrient requirements. Specifically:

1. Genes for which null-type mutants (deletions or transposon insertions) have been isolated are classified into the "nonessential" category.
2. Genes located within the deleted regions of characterized deletion mutants are classified into the "nonessential" category.
3. Genes that do not fall under (1) or (2) and for which conditional lethal mutants have been isolated are classified into the "essential" category.

1.1.2. No Experimental Evidence Is Available

4. Structural genes for ribosomal proteins are classified as "essential," except for those that have been reported to be dispensable.
5. The *argX*, *cysT*, *glyT*, *hisR*, *leuU*, *leuW*, *leuZ*, *proL*, *proM*, *serT*, *serV*, *thrU*, and *trpT* genes encoding unique transfer RNAs (tRNAs) are classified into the "essential" category.
6. The *hisS* and *argS* genes coding for unique aminoacyl-tRNA synthases are classified into the "essential" category.

7. Genes involved in flagellation, motility, and chemotaxis (*flg*, *flh*, *fli*, *mot*, *che*, *tap*, and *tar*) are classified into the "nonessential" category.
8. The *hem* genes are classified as "nonessential" if the corresponding mutants are able to grow in the presence of exogenous porphyrin is medium.

1.1.3. Others

The genes that do not correspond with those listed in **Section 1.1.1** or **Section 1.1.2** are classified into the "unknown" category.

1.2. Minimal Genome and Large-Deletions Data

Identification of the gene products that play an essential role in an organism's functional repertory is important to understanding the mechanism of cell proliferation. However, it is not a simple subject because there are several different criteria that define gene essentiality.

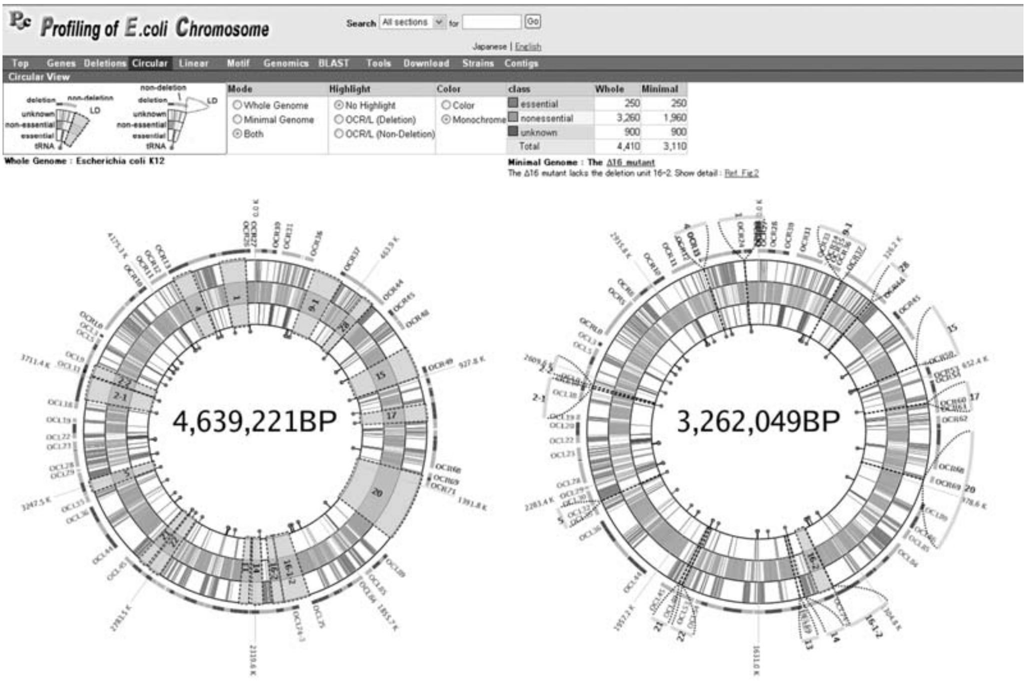
Genome-wide gene knockout mutant collections provide important information, but single-gene mutants are not enough, as they do not account for potential complementation or interactions between genes products. Creating an *E. coli* strain with a minimal genome sufficient to sustain its growth is a challenging approach, but in a sense it may provide the ultimate solution to the problem of gene essentiality. Thus far, *E. coli* genome size has been decreased to about 30% of the parental chromosome, *E. coli* MG1655 genome (*[1]* and **Chapter 18**). The PEC database provides both the original and the minimal genome sizes for each mutant harboring multiple deletions (as shown in **Fig. 2**). Locations of all deletions are indicated in the whole-genome view, and the corresponding primer sequences are also available. We are planning an extensive update of the deletion information in the near future.

1.3. Protein Structural Features

Protein families as well as motifs found in *E. coli* gene products by homology search against Pfam (http://pfam.wustl.edu/) and PROSITE (http://au.expasy.org/prosite/) databases are shown for each gene in a detailed table. Distributions of regions homologous to Pfam families are displayed graphically and the corresponding amino acid sequences are highlighted in the sequence. PEC also provides a summary table of these structural features for all gene products.

1.4. Homologues of E. coli Genes in More than 200 Bacterial Genomes

BLAST searches of *E. coli* genes against other sequenced bacterial genomes with different E-values are carried out in advance, and the results are provided in PEC. At present, a summary table of similarities between each *E. coli* gene and 236 bacterial genomes, including proteobacteria, Firmicutes, actinobacteria, and so forth, is available. This table may help researchers to overview a distribution of orthologous genes among various bacterial genomes without performing a large-scale BLAST search at their ends. We plan to install a more sophisticated viewer as the number of sequenced genomes has increased significantly.



Profiling of E.coli Chromosome

Search All sections for Go

Japanese | English

Mode: ○ Whole Genome ○ Minimal Genome ● Both

Highlight: ● No Highlight ○ OCR/L (Deletion) ○ OCR/L (Non-Deletion)

Color: ○ Color ● Monochrome

class	Whole	Minimal
▨ essential	250	250
▨ nonessential	3,260	1,960
▨ unknown	900	900
Total	4,410	3,110

Whole Genome : Escherichia coli K12

Minimal Genome : The Δ16 mutant

The Δ16 mutant lacks the deletion unit 16-2. Show detail : Ref. Fig.2

Fig. 2. Chromosome maps of *E. coli* MG1655 parental genome (left) and the minimal genome (right). Starting from the inside: gene essentiality is shown on the first ring (essential), the second ring (nonessential), and the third ring (unknown).

1.5. E. coli *Strains and Other Genetic Resources*

The *E. coli* strain collection (http://www.shigen.nig.ac.jp/ecoli/strain/) is supported by the National BioResource Project (**Note 1**). This collection includes:

1. A collection of 2552 *E. coli* mutant (single- and multiple-gene mutants, deletion, and transposon insertion) derived from individual researchers.
2. A complete set of *E. coli* ORFs cloned in vector plasmids *(2)*.
3. The cosmid library of the *E. coli* W3110 chromosome *(3)*.
4. A collection of 2112 hybrid *E. coli* K-12–ColE1 pLC plasmid *(4)*.
5. Genome-wide mutant collection of 3840 in-frame single-gene deletion mutants (the Keio collection (*[5]* and **Chapter 11**).
6. The set of 126 large-scale chromosomal deletion mutants (*[1]* and **Chapter 18**).
7. The collection of 6404 transposon insertion mutants (**Chapter 13**).

PEC also maintains a rich collection of cloning vectors for *E. coli*. All collections are open to the public and available for distribution.

1.6. Genomic Viewers

PEC provides a circular and a linear view for the *E. coli* chromosome. The circular view shows the distribution of genes categorized by essentiality, and each component of the linear view is linked to detailed information. The PEC user can select either a color or a monochrome view. The circular view displays the distribution of essential genes and distribution of long-deletion regions. With the linear view, users can access contig sequences, gene details, and deletion information for a specified region.

1.7. Files Available for Download

PDF files of linear view in three different scales and a tab-separated text file containing all PEC genes are downloadable from the PEC site. GenomePaint, a stand-alone map-drawing tool, is also available from the same site. GenomePaint creates image files of the PEC-type circular and linear genome views from text files.

Note

1. The National BioResource Project (NBRP) aims to enable Japan to structurally provide systematic accumulation, storage, and provision of nationally recognized bioresources that are used widely in life science research, including experimental animals, plants, cells, and DNA and other genetic materials from a variety of species. The NBRP has set up a central resource center in charge of providing a framework for collection, storage, and distribution of such materials for each important genetic model organism. This project started in July 2002 as part of the "Research Revolution 2002 (RR2002)" program of the Ministry of Education, Culture, Sports, Science, and Technology. At present (July 2006), 24 central resource centers including NBRP–*E. coli* and the information center (http://www.nbrp.jp/) are involved in this project.

Acknowledgments

The *E. coli* Genetic Resource Committee of Japan supports the construction and maintenance of the PEC database.

References

1. Hashimoto, M., Ichimura, T., Mizoguchi, H., Tanaka, K., Fujimitsu, K., Keyamura, K., et al. (2005) Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. *Mol. Microbiol.* **55**, 137–149.
2. Kitagawa, M., Ara, T., Arifuzzaman, M., Ioka-Nakamichi, T., Inamoto, E., Toyonaga, H., and Mori, H. (2005) Complete set of ORF clones of *Escherichia coli* ASKA library (A complete set of *E. coli* K-12 ORF archive): Unique Resources for Biological Research. *DNA Res.* **12**, 291–299.
3. Tabata, S., Higashitani, A., Takanami, M., Akiyama, K., Kohara, Y., Nishimura, Y., et al. (1989) Construction of an ordered cosmid collection of the *Escherichia coli* K-12 W3110 chromosome. *J. Bacteriol.* **171**, 1214–1218.
4. Nishimura, A., Akiyama, K., Kohara, Y., and Horiuchi, K. (1992) Correlation of a subset of the pLC plasmids to the physical map of *Escherichia coli* K-12. *Microbiol. Rev.* **56**, 137–151.
5. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., et al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Systems Biol.* **2**, 2006. 008.

27

Gene Essentiality Analysis Based on DEG, a Database of Essential Genes

Chun-Ting Zhang and Ren Zhang

Summary

Essential genes are the genes that are indispensable for the survival of an organism. The genome-scale identification of essential genes has been performed in various organisms, and we consequently constructed DEG, a Database that contains currently available essential genes. Here we analyzed functional distributions of essential genes in DEG, and found that some essential-gene functions are even conserved between the prokaryote (bacteria) and the eukaryote (yeast), e.g., genes involved in information storage and processing are overrepresented, whereas those involved in metabolism are underrepresented in essential genes compared with non-essential ones. In bacteria, species specificity in functional distribution of essential genes is mainly due to those involved in cellular processes. Furthermore, within the category of information storage and processing, function of translation, ribosomal structure, and biogenesis are predominant in essential genes. Finally, some potential pitfalls for analyzing gene essentiality based on DEG are discussed.

Key Words: COG; database; DEG; essential gene.

1. Introduction

The complete sequencing of a large number of genomes has revolutionized biomedical research. The sequencing of the first bacterial genome, the genome of *Haemophilus influenzae*, was finished in 1995 *(1)*. Eleven years later, more than 300 bacterial genome sequences have been deposited in public databases, such as GenBank, EMBL, and DDBJ. In addition, many eukaryotic genomes have also been sequenced; for example, the human genome *(2)*, which contains 3 × 10^9 nucleotides. This vast amount of sequence information has fundamentally influenced biomedical research in almost every field.

One piece of information that can be gained from the complete genome sequence of an organism is that all the gene and protein sequences of the sequenced organism can be revealed. A natural question is which genes are indispensable for the survival of this organism and which genes are dispensable. Essential genes are the genes that are

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

absolutely required for the survival of an organism under certain conditions, such as in rich medium. The essential genes from an organism form a minimal gene set for this organism. An ambitious idea is to make an autonomous cell based on the minimal gene set *(3–6)*. This is a fundamental question in biology because genes in the minimal gene set encode the most basic functions for one particular organism and are even required for other organisms.

The elucidation of the minimal gene set is not only critical to provide insights in understanding cellular functions but also has many important applications. One application is that essential genes are good candidates for antibacterial drugs because most drugs target the genes involved in critical cellular processes. Indeed, some antibiotics have been designed based on this principle. For example, DNA gyrase is an essential prokaryotic enzyme that catalyzes chromosomal DNA supercoiling. Ciprofloxacin, a new fluoroquinolone, is a potent, broad-spectrum antibacterial agent that blocks bacterial DNA replication by inhibiting DNA gyrase *(7)*.

The genome-scale identification of essential genes is made possible with the advent of completed genome sequences and large-scale gene-inactivation technologies, such as targeted gene inactivation, transposon-based mutagenesis, and genetic footprinting. In addition, some bioinformatics tools were also developed for essential gene identification *(8, 9)*. Sometimes, bioinformatics tools were used together with experimental methods; for example, 113 essential genes of *Streptococcus pneumoniae* were identified by bioinformatics analysis followed by targeted gene disruption *(10)*.

2. Construction of a Database of Essential Genes

Today, genome-scale essential gene identification has been performed in a number of organisms. For example, essential genes have been identified in *Bacillus subtilis* by insertional mutagenesis *(11)*, *Escherichia coli* by genetic footprinting *(12)*, *Haemophilus influenzae* by high-density transposon mutagenesis *(13)*, *Mycoplasma genitalium* by transposon-based mutagenesis *(4)*, *Staphylococcus aureus* by antisense RNA technique *(14, 15)*, *Streptococcus pneumoniae* Rx-1 by bioinformatics analysis followed by targeted gene disruption *(10)*, *Vibrio cholerae* by transposon-based mutagenesis *(16)*, and in *Saccharomyces cerevisiae* yeast by systematic gene inactivation *(17)* (**Table 1**).

We have constructed a database of essential genes (DEG) *(18)*, which includes the identified essential genes in the genomes of *M. genitalium, H. influenzae*, *V. cholerae*, *S. aureus*, *E. coli*, and yeast (**Note 1**). The essential genes in the yeast genome were extracted from the yeast genome database (http://www.mips.biochem.mpg.de/proj/yeast), which is maintained by the Munich Information Center for Protein Sequences *(19)*.

Each entry has a unique DEG identification number, gene reference number, gene function, and sequence. All information is stored and operated by using an open-source database management system, MySQL. Users can browse and extract all the records of these entries. In addition, users can also search for essential genes in DEG by gene functions or names. Furthermore, we have installed the BLAST program locally. Therefore, users can perform BLAST searches for query sequences against all essential genes in DEG (**Note 2**). DEG is freely available at http://tubic.tju.edu.cn/deg.

Table 1
Essential Genes of Various Organisms Currently Available in DEG 1.0

	Organism	No. of essential genes	Total gene no.	% of essential genes	Genome length (bp)	Method	References
1	*Bacillus subtilis*	271	4234	6.61	4,214,814	Insertional mutagenesis based on designed vectors	(*11*)
2	*Escherichia coli*	620	3746	16.55	4,639,221	Genetic footprinting	(*12*)
3	*Haemophilus influenzae*	638	1788	35.68	1,830,138	High-density transposon mutagenesis	(*13*)
4	*Mycoplasma genitalium*	265–300	517	51.26–58.03	580,074	Transposon-based mutagenesis	(*4*)
5	*Staphylococcus aureus*	658	2695	24.42	2,814,816	Rapid shotgun antisense RNA	(*14, 15*)
6	*Streptococcus pneumoniae* Rx-1	113	2306	32.56	2,160,837	Bioinformatics analysis followed by targeted gene disruption	(*10*)
7	*Vibrio cholerae*	5	4007	1.25	4,033,464	Transposon-based mutagenesis	(*16*)
8	*Saccharomyces cerevisiae*	878	5885	14.92	12,068,000	Systematic gene inactivation	(*17*)

3. Analysis of Functional Class Distributions of Essential Genes Based on DEG

Essential genes are genes that are indispensable to supporting cellular life. Therefore, it is of considerable interest to investigate the functional classes of the available essential genes. With the availability of essential genes from multiple genomes, it is possible to investigate the functional conservation of essential genes across organisms.

Based on homologous relationships, all conserved genes are classified. Consequently, 2791 clusters of orthologous groups (COGs) have been delineated *(20)*. The genes within the same cluster usually have the same function; therefore, the COGs compose a framework for functional analysis of genomes. We used the COG classification to analyze the functional distribution of essential and nonessential genes (**Note 3**). Among the available essential gene studies, we chose only the data sets generated by methods that meet the following requirements: (1) the method identifies essential genes experimentally and not by bioinformatics analysis; (2) the method systematically identifies essential genes on a whole-genome scale; and (3) the COG classification of genes in a genome is available in the COG database. Consequently, we performed the analysis for the genomes of *B. subtilis, H. influenzae, M. genitalium*, and yeast.

According to COG classification (http://www.ncbi.nlm.nih.gov/COG/), genes are classified into four broad functional categories; that is, information storage and processing, cellular processes, metabolism, and poorly characterized. To have a picture of the global functional distribution of essential and nonessential genes in bacteria, the three bacterial essential and nonessential genes were pooled based on the four broad functional categories (**Fig. 1A**). Of essential genes, the four functional categories (i.e.,

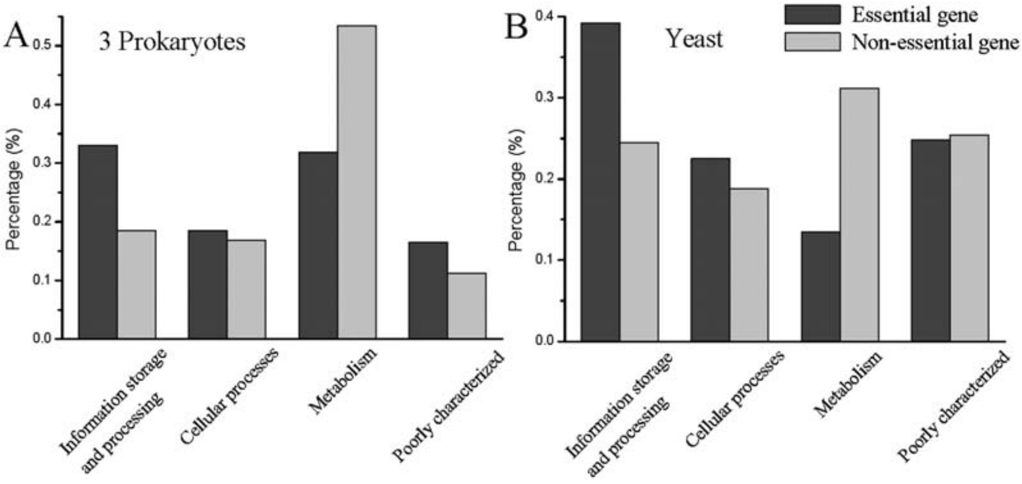


Fig. 1. Distribution of the four broad functional categories of essential and nonessential genes in **(A)** three bacteria (*B. subtilis, H. influenzae*, and *M. genitalium*) and **(B)** yeast. Some functional category distributions are highly conserved between bacteria and yeast. For instance, in bacteria, the category of information storage and processing is overrepresented in essential genes compared with that of nonessential genes, whereas the category of metabolism is underrepresented in essential genes. This pattern also holds for yeast genes. Refer to **Section 3** for details.

information storage and processing, cellular processes, metabolism, and poorly characterized) comprise 33.1%, 18.5%, 31.9%, and 16.5%, respectively, whereas the four functional categories of nonessential genes comprise 18.5%, 16.9%, 53.4%, and 11.2%, respectively. Therefore, the categories of information storage and processing and metabolism comprise more than two-thirds of essential genes, whereas in nonessential genes, more than one-half of genes are involved in metabolism. Therefore, one difference between essential and nonessential genes is that the category of information storage and processing is overrepresented in essential genes (33.1%) compared with that of nonessential genes (18.5%). In contrast, the category of metabolism is underrepresented in essential genes (31.9%) compared with that of nonessential genes (53.4%). The proportions of genes involved in cellular processes are similar, that is, 18.5% and 16.9%, respectively, for essential and nonessential genes.

We also performed a similar analysis for yeast genes (**Fig. 1B**). Although yeast is a eukaryotic organism, it is striking that the distribution of functional categories among essential and nonessential genes is highly similar to that of bacterial genes. The four functional categories of yeast essential genes comprise 39.5%, 22.5%, 13.5%, and 24.8%, respectively, whereas those of nonessential genes are 24.5%, 18.8%, 31.2%, and 25.4%, respectively. Therefore, the observations of functional distribution of bacterial essential genes still hold in the case of yeast. That is, the category of information storage and processing is overrepresented in essential genes (39.5%) compared with that of nonessential genes (24.5%), and the category of metabolism is underrepresented in essential genes (13.5%) compared with that of nonessential genes (31.2%). The conservation of the functional category distribution between prokaryotes (the three bacteria) and the eukaryote (yeast) suggests that some functions of essential genes are likely to be universal and required for organisms in different kingdoms. In addition, a large proportion of essential genes are poorly characterized, that is, either classified as general function prediction or function unknown, suggesting that many novel gene functions that are critical to support cellular life are still elusive.

We next examined the distribution of these four functional categories in each of the three bacterial genomes (**Fig. 2**). Many features of the distribution are well conserved among the three bacteria, whereas some are different. In all three bacteria, genes involved in information storage and processing represent a larger proportion in essential genes than that of nonessential genes. The proportions of the category of information storage and processing in essential and nonessential genes in *B. subtilis* are 48.7% and 21.1%; in *M. genitalium*, 44.7% and 27.5%; and in *H. influenzae*, 25.0% and 11.5%, respectively. Furthermore, in all three bacteria, genes involved in metabolism represent a smaller proportion in essential genes than that of nonessential genes. The proportions of the category of metabolism in essential and nonessential genes in *B. subtilis* are 26.3% and 47.6%; in *M. genitalium*, 26.5% and 34.8%; and in *H. influenzae*, 36.8% and 71.6%, respectively. However, some patterns of distribution are indeed different. For instance, proportions of genes involved in cellular processes in *B. subtilis* are similar between essential and nonessential genes (i.e., 18.4% and 18.1%, respectively). In *M. genitalium*, the proportion of genes involved in cellular processes of essential genes is less than that of nonessential genes (14.6% and 18.8%, respectively), whereas in *H. influenzae*, the proportion of essential genes is more than that of nonessential

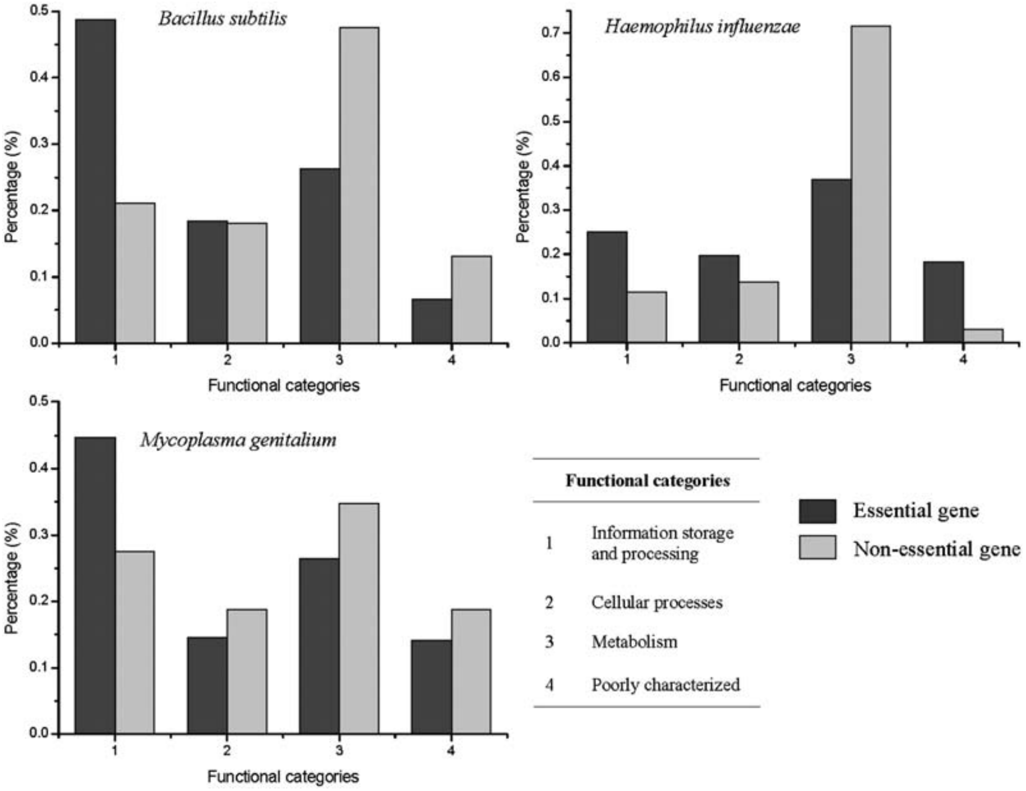
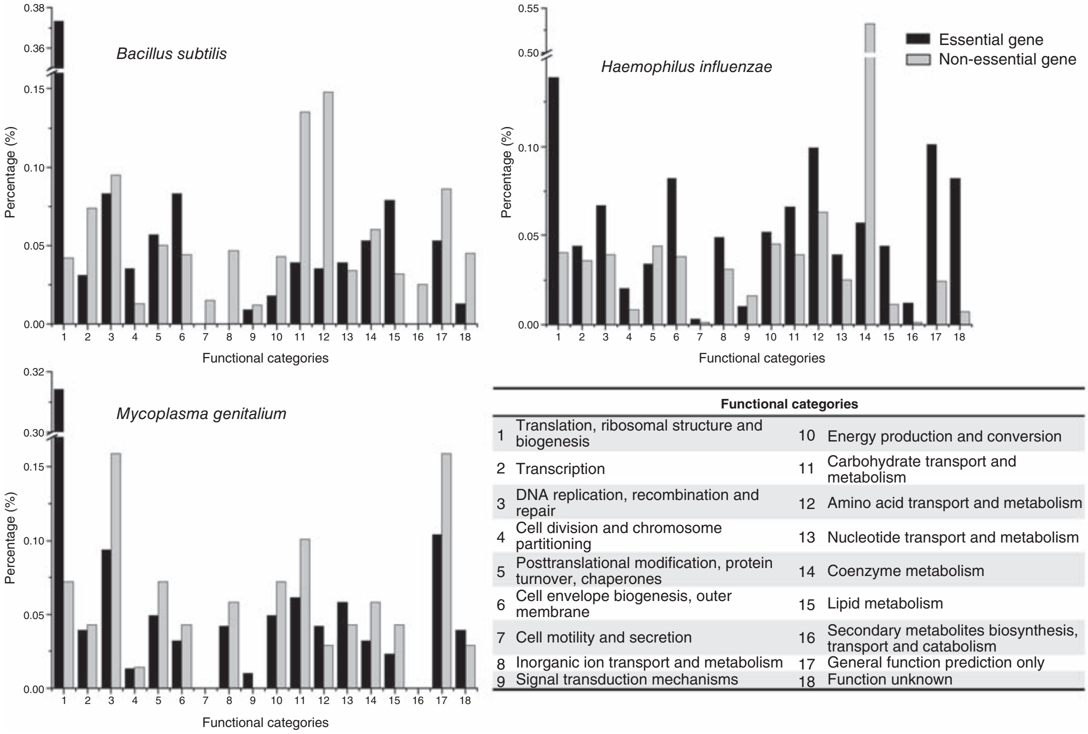


Fig. 2. Distribution of the four broad functional categories of essential and nonessential genes in three bacteria (*B. subtilis, H. influenzae*, and *M. genitalium*). In all three bacteria, genes involved in information storage and processing represent a larger proportion in essential genes than that of nonessential genes. In contrast, in all three bacteria, genes involved in metabolism represent a smaller proportion in essential genes than that of nonessential genes. Conserved distributions of functional classes suggest that functions of these essential genes are required for all three bacteria, whereas the different distribution patterns suggest that essential genes have some kind of species specificity. Based on the analysis of these three bacteria, it is possible that the essential genes involved in information storage and processing are well conserved between species, whereas the species specificity is mainly due to the essential genes involved in cellular processes. Refer to **Section 3** for details.

genes (19.8% and 13.7%, respectively). Conserved distributions of functional categories suggest that the functions of these essential genes are required for all three bacteria, whereas the different distribution patterns suggest that essential genes have some kind of species specificity. Based on the analysis of these three bacteria, it is possible that the essential genes involved in information storage and processing are well conserved between species, whereas the species specificity is mainly due to the essential genes involved in cellular processes.

In the COG database, each broad category is composed of some functional classes. The category of information storage and processing is further classified into (1) translation, ribosomal structure, and biogenesis; (2) transcription; and (3) DNA replication, recombination, and repair. The category of cellular processes is further classified into (1) cell division and chromosome partitioning; (2) posttranslational modification, protein turnover, and chaperones; (3) cell envelope biogenesis and outer membrane; (4) cell motility and secretion; (5) inorganic ion transport and metabolism; and (6) signal transduction mechanisms. The category of metabolism is further classified into (1) energy production and conversion; (2) carbohydrate transport and metabolism; (3) amino acid transport and metabolism; (4) nucleotide transport and metabolism; (5) coenzyme metabolism; (6) lipid metabolism; and (7) secondary metabolite biosynthesis, transport, and catabolism. The category of poorly characterized is further classified into (1) general function prediction only and (2) function unknown. According to these 18 functional classes, we examined the functional distribution of essential genes among the three bacterial genomes (**Fig. 3**).

The essential genes involved in translation, ribosomal structure, and biogenesis are predominant in all three bacteria. The proportion of such genes in *B. subtilis* is 37.3%; in *M. genitalium*, 31.4%; and in *H. influenzae*, 13.9%. The proportions of such class in nonessential genes in these three bacteria are 4.2%, 7.2%, and 4.0%, respectively. Therefore, in all three bacteria, the proportion of genes involved in translation, ribosomal structure, and biogenesis is more than that in nonessential genes. Within the class of translation, ribosomal structure, and biogenesis, the distribution of genes is also



Functional categories			
1	Translation, ribosomal structure and biogenesis	10	Energy production and conversion
2	Transcription	11	Carbohydrate transport and metabolism
3	DNA replication, recombination and repair	12	Amino acid transport and metabolism
4	Cell division and chromosome partitioning	13	Nucleotide transport and metabolism
5	Posttranslational modification, protein turnover, chaperones	14	Coenzyme metabolism
6	Cell envelope biogenesis, outer membrane	15	Lipid metabolism
7	Cell motility and secretion	16	Secondary metabolites biosynthesis, transport and catabolism
8	Inorganic ion transport and metabolism	17	General function prediction only
9	Signal transduction mechanisms	18	Function unknown

Fig. 3. Distribution of 18 functional classes of essential and nonessential genes among three bacteria (*B. subtilis, H. influenzae*, and *M. genitalium*). Refer to **Section 3** for details.

strongly biased. Most of the essential genes in this functional class encode ribosomal proteins. For instance, in *B. subtilis*, genes encoding ribosomal proteins comprise 61.2% of essential genes in this functional class. In contrast, there are also differences of the functional distribution among both essential and nonessential genes, which may reflect species specificity. For instance, in the genomes of *H. influenzae*, genes involved in coenzyme metabolism themselves comprise more than half of nonessential genes (53.2%), whereas in the genomes of *B. subtilis* and *M. genitalium*, they only comprise 6.0% and 5.8%, respectively.

4. Gene Essentiality Predictions

The *in silico* identification of essential genes can probably be traced back to a prediction of minimal gene set soon after two completed genome sequences became available in 1995 ***(8)***. The algorithm used in this work was that genes that are conserved between two evolutionarily distant bacteria are likely to be essential. *M. genitalium* has an extremely small genome (i.e., 0.58 megabase) and has only 468 protein-coding genes. It was claimed that *M. genitalium* has a minimal gene complement ***(4)***. *H. influenzae* also has a relatively small genome, 1.83 megabase, and has 1700 protein-coding genes ***(1)***. Importantly, *M. genitalium* is Gram-positive, whereas *H. influenzae* is Gram-negative, and they are likely to be separated from their last common ancestor by at least 1.5 billion years of evolution ***(21)***. Therefore, genes shared by the two genomes are likely to be essential. About 250 orthologs were found, which were predicted to be essential genes ***(8)***. Indeed, many of the predicted essential genes were validated in future experiments ***(6)***.

Another algorithm proposed later was based on this notion: genes that are conserved across different bacterial species are likely to be essential ***(9)***. Therefore, based on a Web tool, users can determine concordances of putative gene products that show sets of proteins conserved across one set of user-specified genomes and are not present in another set of user-specified genomes. Based on this method, the authors were able to identify a known target of quinolone antibiotics ***(9)***.

Both of the above methods are based on the identification of conserved genes among two or more genomes. At the time when these two methods were proposed, not much information on essential genes that are identified based on experimental evidence was available. Today, many genome-scale gene inactivation experiments have been performed, and therefore, many essential genes have been identified. DEG contains the record of those experimentally determined essential genes. We propose a new method that determines the gene essentiality based on homologous search against DEG. Because the functions encoded by essential genes are indispensable to supporting cellular life, these functions can be considered a foundation of life itself. Indeed, based on the analysis of functional class distributions of essential genes presented in the previous section, some functions of essential genes are well conserved among different organisms. It is even believed that some basic functions and principles are common to all cellular life on this planet ***(22)***. Therefore, based on BLAST searches, if the queried sequences have homologous genes in DEG, it is likely that the queried genes are also essential. In addition, by performing BLAST searches against DEG (**Note 2**) for all the protein-coding genes in a genome, it is possible to define the putative essential genes for proteomes of the newly sequenced genomes.

Notes

1. It is helpful to keep in mind that the essential genes listed in DEG, although validated by experiments, are essential only under certain conditions such as in rich or minimal media. These laboratory conditions may be quite different from those of a real environment or may not even exist in a real environment.
2. Caution must be taken to interpret the results of a BLAST search against DEG. Because DEG contains essential gene records from many essential gene projects, the essential genes predicted based on a homologous search against DEG are likely to represent an overprediction (i.e., false-positive prediction rates are likely to be higher than false-negative prediction rates). Some essential genes are essential for one species but not necessarily essential for another species. In other words, *bona fide* essential genes are not likely to be missed by this method, but some predicted essential genes may not be essential due to species specificity.
3. While studying the functional distribution of nonessential genes, such as in **Figure 1**, we defined the nonessential genes to be the genes that exclude essential genes. However, caution must be taken because some nonessential genes so defined are not necessarily nonessential. Because most essential gene projects assess viability of an organism based on technologies that inactivate single genes, and if two essential genes have redundant functions (e.g., two homologous essential genes), the inactivation of one gene may not affect the viability of the organism. Therefore, some nonessential genes may in fact encode products that are essential for the survival of an organism.

Acknowledgments

This study was supported in part by the 973 Project of China (grant 1999075606).

References

1. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.
2. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
3. Cho, M. K., Magnus, D., Caplan, A. L., and McGee, D. (1999) Policy forum: genetics. Ethical considerations in synthesizing a minimal genome. *Science* **286**, 2087, 2089–2090.
4. Hutchison, C. A., Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., et al. (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286**, 2165–2169.
5. Mushegian, A. (1999) The minimal genome concept. *Curr. Opin. Genet. Dev.* **9**, 709–714.
6. Koonin, E. V. (2000) How many genes can make a cell: the minimal-gene-set concept. *Annu. Rev. Genomics Hum. Genet.* **1**, 99–116.
7. Fisher, L. M., Lawrence, J. M., Josty, I. C., Hopewell, R., Margerrison, E. E., and Cullen, M. E. (1989) Ciprofloxacin and the fluoroquinolones. New concepts on the mechanism of action and resistance. *Am. J. Med.* **87**, 2S–8S.
8. Mushegian, A. R., and Koonin, E. V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10268–10273.

9. Bruccoleri, R. E., Dougherty, T. J., and Davison, D. B. (1998) Concordance analysis of microbial genomes. *Nucleic Acids Res.* **26**, 4482–4486.
10. Thanassi, J. A., Hartman-Neumann, S. L., Dougherty, T. J., Dougherty, B. A., and Pucci, M. J. (2002) Identification of 113 conserved essential genes using a high-throughput gene disruption system in *Streptococcus pneumonia*e. *Nucleic Acids Res.* **30**, 3152–3162.
11. Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., Arnaud, M., et al. (2003) Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4678–4683.
12. Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balazsi, G., Ravasz, E., Daugherty, M. D., et al. (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684.
13. Akerley, B. J., Rubin, E. J., Novick, V. L., Amaya, K., Judson, N., and Mekalanos, J. J. (2002) A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 966–971.
14. Ji, Y., Zhang, B., Van, S. F., Horn, W. P., Woodnutt, G., Burnham, M. K., and Rosenberg, M. (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science* **293**, 2266–2269.
15. Forsyth, R. A., Haselbeck, R. J., Ohlsen, K. L., Yamamoto, R. T., Xu, H., Trawick, J. D., et al. (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol. Microbiol.* **43**, 1387–1400.
16. Judson, N., and Mekalanos, J. J. (2000) TnAraOut, a transposon-based approach to identify and characterize essential bacterial genes. *Nat. Biotechnol.* **18**, 740–745.
17. Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391.
18. Zhang, R., Ou, H. Y., and Zhang, C. T. (2004) DEG: a database of essential genes. *Nucleic Acids Res.* **32** (Database issue), D271–D272.
19. Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., et al. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34.
20. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**, 22–28.
21. Doolittle, R. F., Feng, D. F., Tsang, S., Cho, G., and Little, E. (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* **271**, 470–477.
22. Peterson, S. N., and Fraser, C. M. (2001) The complexity of simplicity. *Genome Biol.* **2**, comment 2002.1–2002.8.

28

Detection of Essential Genes in *Streptococcus pneumoniae* Using Bioinformatics and Allelic Replacement Mutagenesis

Jae-Hoon Song and Kwan Soo Ko

1. Introduction

Although the emergence and spread of antimicrobial resistance in major bacterial pathogens for the past decades poses a growing challenge to public health, discovery of novel antimicrobial agents from natural products or modification of existing antibiotics cannot circumvent the problem of antimicrobial resistance. The recent development of bacterial genomics and the availability of genome sequences allow the identification of potentially novel antimicrobial agents. The cellular targets of new antimicrobial agents must be essential for the growth, replication, or survival of the bacterium. Conserved genes among different bacterial genomes often turn out to be essential ***(1, 2)***. Thus, the combination of comparative genomics and the gene knock-out procedure can provide effective ways to identify the essential genes of bacterial pathogens ***(3)***. Identification of essential genes in bacteria may be utilized for the development of new antimicrobial agents because common essential genes in diverse pathogens could constitute novel targets for broad-spectrum antimicrobial agents.

In this chapter, we introduce a rapid and efficient method for the identification of essential genes in *Streptococcus pneumoniae* that combines comparative genomics and allelic replacement mutagenesis.

2. Materials

2.1. Streptococcus pneumoniae

1. *S. pneumoniae* strain D39.
2. Todd-Hewitt broth or agar (Difco, Becton-Dickinson, Sparks, MD) supplemented with 0.5% yeast extract (Difco) (THYE).

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

3. Kanamycin (Sigma-Aldrich, St. Louis, MO).
4. Blood agar plate (BAP).
5. Resuspending solution: TE buffer (10 mM Tris-HCl, pH 8.0, and 1 mM EDTA), 0.005% sodium deoxycholate, and 0.01% SDS.
6. Proteinase K (Sigma-Aldrich).
7. Phenol/chloroform/isoamyl alcohol (25:24:1) (Invitrogen, Carlsbad, CA).
8. Oligonucleotide primers: Kan-F (5′-AAC AGT GAA TTG GAG TTC GTC TTG TTA TA-3′), Kan-R (5′-GCT TTT TAG ACA TCT AAA TCT AGG TA-3′), and others.
9. Agarose electrophoresis and polymerase chain reaction (PCR) equipment.
10. CoreOne PCR purification kit (CoreBioSystem, Seoul, Korea).
11. Competence medium: THYE, 0.2% bovine serum albumin, 0.01% $CaCl_2$, and 100 ng/mL peptide pheromone CSP (Takara Korea, Seoul, Korea): H-Glu-Met-Arg-Leu-Ser-Lys-Phe-Arg-Asp-Phe-Ile-Leu-Gln-Arg-Lys-Lys-Oh.

3. Methods

3.1. Selection of Target Genes by Bioinformatics

S. pneumoniae R6 genome sequence data are used for selection of target genes. Target genes are selected using the Microbial Concordance Tool (*4, 5*) as follows: the amino acid sequences of 2046 *S. pneumoniae* R6 open reading frames (ORFs) are compared with those of *Bacillus subtilis*, *Enterococcus faecalis*, *Escherichia coli*, and *Staphylococcus aureus*, and genes of more than 40% amino acid sequence identity to the corresponding genes in at least two of the other species are selected.

3.2. Preparation of Competent Cells

1. To prepare competent cells, *S. pneumoniae* is plated and cultured on a fresh blood agar (**Note 1**).
2. One colony is picked from a cultured plate and resuspended in 1.5 mL THYE. One hundred microliters of the resuspension is used to inoculate 50 mL of the same medium, which is grown at 37°C overnight.
3. Five milliliters of the culture is added to 45 mL fresh medium and is grown at 37°C to OD_{600} for 4 to 5 h.
4. Sterile glycerol is added to a final concentration of 10%, and cells are aliquoted in 1-mL samples, frozen in a dry ice–ethanol bath, and stored at −80°C.

3.3. Allelic Replacement Mutagenesis

3.3.1. Extraction of Genomic DNA

1. *S. pneumoniae* D39 is grown overnight on a blood agar plate at 37°C in 5% CO_2 for extraction of genomic DNA.
2. A single colony is removed with an inoculating loop and resuspended in 20 mL Todd-Hewitt agar supplemented with 0.5% yeast extract (THYE) with 400 μg/mL sterile sodium bicarbonate.

3. The bacterial cells are grown at 37°C until an OD_{600nm} reaches 0.4 to 0.6 and are then chilled on ice and harvested by centrifugation at 5000 rpm for 15 min at 4°C.
4. The pellet is resuspended and washed once with 20 mL ice-cold TE buffer, centrifuged as above, and the resulting pellet quick frozen at –20°C.
5. The cells are thawed and resuspended in 5 mL TE buffer, and 0.005% sodium deoxycholate and 0.01% SDS added. Cells are lysed by incubation at 37°C for 10 min.
6. After cell lysis, 500 µg/mL proteinase K is added and additionally incubated for 10 min. The cell lysate is gently extracted with an equal volume of phenol/chloroform/isoamyl alcohol (25 : 24 : 1).
7. After centrifugation at 8000 rpm for 10 min, the upper layer is removed and extracted twice with an equal volume chloroform/isoamyl alcohol (24 : 1).
8. The final aqueous extract is brought up to 0.3 M sodium acetate and is overlaid with 2.2 volumes of ethanol. The DNA is spooled onto a glass rod and redissolved in 2 mL TE buffer overnight at 4°C.
9. This preparation is dialyzed against 400 volumes of TE buffer before storage at 4°C. DNA concentration is determined by absorbance at 260 nm and adjusted to 0.5 µg/µL.

3.3.2. Preparation of Kanamycin-Resistance Cassette

A kanamycin-resistance cassette (904 bp) containing KanR from *Staphylococcus aureus* ATCC43300 is amplified using the primer set, Kan-F (5′-AAC AGT GAA TTG GAG TTC GTC TTG TTA TA-3′) and Kan-R (5′-GCT TTT TAG ACA TCT AAA TCT AGG TA-3′) (*6, 7*).

3.3.3. Two-Step PCR Protocol

For allelic replacement mutagenesis, two-step PCR is performed (**Fig. 1**). Two pairs of gene-specific primers, L-F/L-R and R-F/R-R, are used to amplify the left and right flanking regions of each target gene, generating PCR products of 500 to 800 bp in length. Primers L-R and R-F consist of 21 nucleotides (5′-GAC GAA CTC CAA TTC ACT GTT-3′ and 5′-AGA TTT AGA TGT CTA AAA AGC-3′, respectively), which are identical to the promoter region and the 3′-end of the Kan^R gene, plus 23 nucleotides of target gene–specific sequence.

PCR amplifications are run in 96-well format under the following conditions: 30 cycles of 94°C for 1 min, 55°C for 1 min, and 72°C for 1 min 30 s, and final extension of 72°C for 10 min. Each PCR product is purified using Core-One PCR purification kit (CoreBioSystem). A template mixture of the amplified Kan^R gene and two PCR products flanking the target gene are then subjected to a second PCR amplification to produce a linear fused product using primers L-F and R-R. The second PCR reaction mix contains in a total volume of 50 µL: 2 µL of each, left and right flanking PCR products and the Kan^R gene cassette, 5 µL of 10× buffer, 1 µL of each primer (L-F and R-R) (25 pmol/µL), 5 µL of dNTP mix (25 mM each), and 1 unit of *Taq* polymerase. The cycling condition are as follows: 30 cycles of 94°C for 40 s, 50°C for 40 s, and 72°C for 2 min 30 s, and the final extension of 72°C for 10 min (**Note 2**).

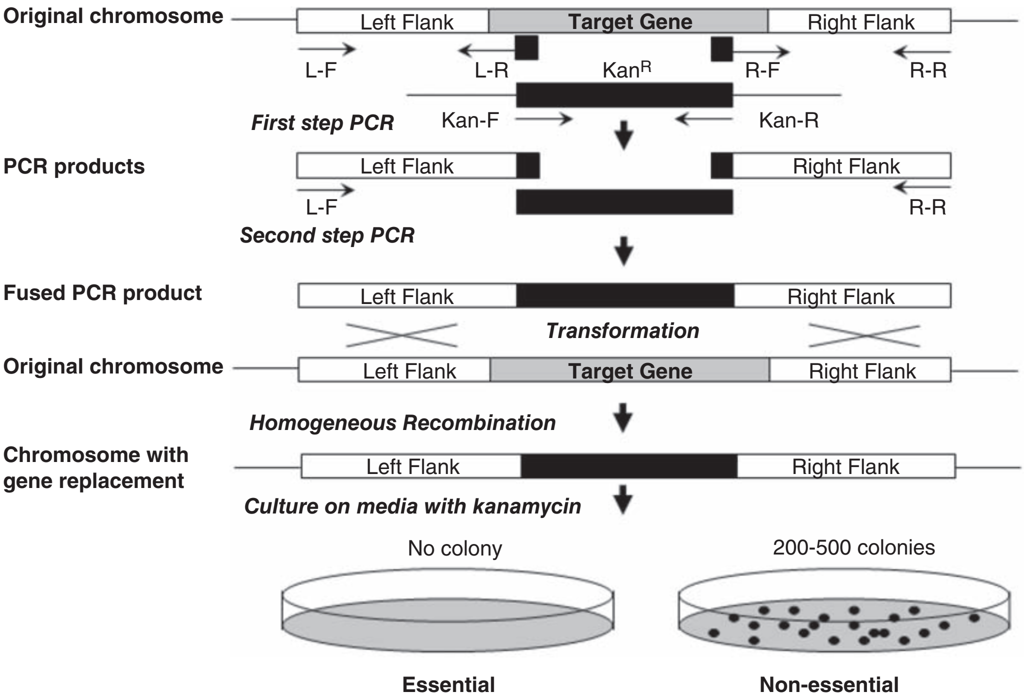


Fig. 1. Two-step PCR procedure to generate a fusion between kanamycin-resistance cassette (Kan^R) and the flanking regions of a target gene: introduction of a fused PCR product into *S. pneumoniae* chromosome via transformation and homologous recombination with the target gene.

3.3.4. *S. pneumoniae* Transformation

The linear fused product produced by the two-step PCR procedure is introduced into the chromosomal genome of *S. pneumoniae* D39 by transformation and homologous recombination (**Fig. 1**; *see* **Note 3**). Pneumococcal transformation is performed as follows *(8)*:

1. 1 µg DNA and 200 µL *S. pneumoniae* D39 competent cells are diluted 1:10 in competence medium containing peptide pheromone CSP (Takara Korea; *see* **Note 4** and *[8]*).
2. Cells are incubated at 37°C for 2.5 to 3 h without shaking and are plated on THYE with 400 µg/mL kanamycin (**Note 5**).
3. Plates are incubated at 37°C for 24 h in a CO_2 incubator.

As a result of introduction of the fused PCR product into the genome of *S. pneumoniae*, the Kan^R gene cassette replaces the chromosomal copy of the target gene, thereby creating a gene knockout. In all transformation experiments, THYE with 5% lysed sheep blood is used for growth of bacterial cells and preparation of competent cells.

3.3.5. Identification of Essential Genes

Typically, inactivation of a nonessential gene produces 300 to 500 Kan^R transformants. If no Kan^R colonies are obtained, the transformation is repeated at least two more times. Genes are regarded as essential if no colonies are observed in all three transformations. If one or more Kan^R colonies are obtained, the target gene is considered nonessential.

3.4. Confirmation of Gene Replacement Events

Targeted gene replacement events are confirmed by PCR assay. Genomic DNA from mutant and wild-type strains are used as PCR templates along with primers L-F and R-R to verify the correct incorporation of a gene replacement construct into the genome. PCR reactions are carried out under the same conditions as described in **Section 3.2.3** (30 cycles of 95°C for 40 s, 50°C for 40 s, and 72°C for 2 min 30 s). The correct incorporation of a fused construct results in larger or smaller PCR product obtained for a mutant strain compared with that for the wild-type strain (**Fig. 2**).

3.5. Evaluation of Potential Polar Effects

The spr0004 and spr0005 genes (SP0004 and SP0005 in TIGR4) have been previously reported as essential (***5***). However, spr0004 was identified as nonessential by allelic replacement mutagenesis using Kan^R cassette without polarity (**Fig. 3**; *see* **Note 6**). Our data suggest that this allelic replacement method can effectively determine essentiality of monocistronic as well as polycistronic ORFs in *S. pneumoniae*.

3.6. Advantages of Allelic Replacement Mutagenesis Coupled with Comparative Genomics

Various methods have been proposed and performed to identify essential genes in several bacterial species (***2, 5, 9–13***). Compared with other techniques, our method has some advantages in identifying essential genes in pathogenic bacteria, including *S. pneumoniae* (***14***). First, stepwise filtering of ORFs through cross-genome comparison with other species based on simple criteria can effectively reduce the number of genes to be tested, as indicated in previous studies (***4, 5***). Second, *a priori* knowledge of target genes makes it unnecessary to sequence *a posteriori* for mutant identification. Third, allelic replacement mutagenesis by two-step PCR does not require cloning into a vector

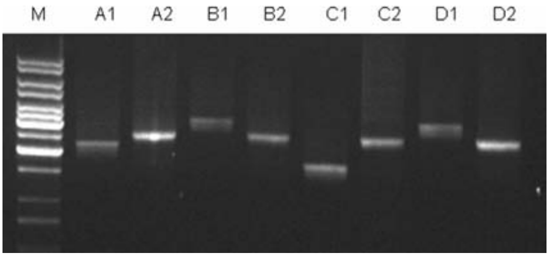


Fig. 2. Confirmation of gene replacement events. Lanes A1, B1, C1, and D1: PCR amplification of the wild-type genes spr0147, spr0232, spr0746, and spr1153 respectively. Lanes A2, B2, C2, and D2: PCR amplification of the corresponding mutant alleles. Lane M: 100-bp ladder marker. Note size differences between PCR products yielded by mutagenized and the corresponding wild-type genes.

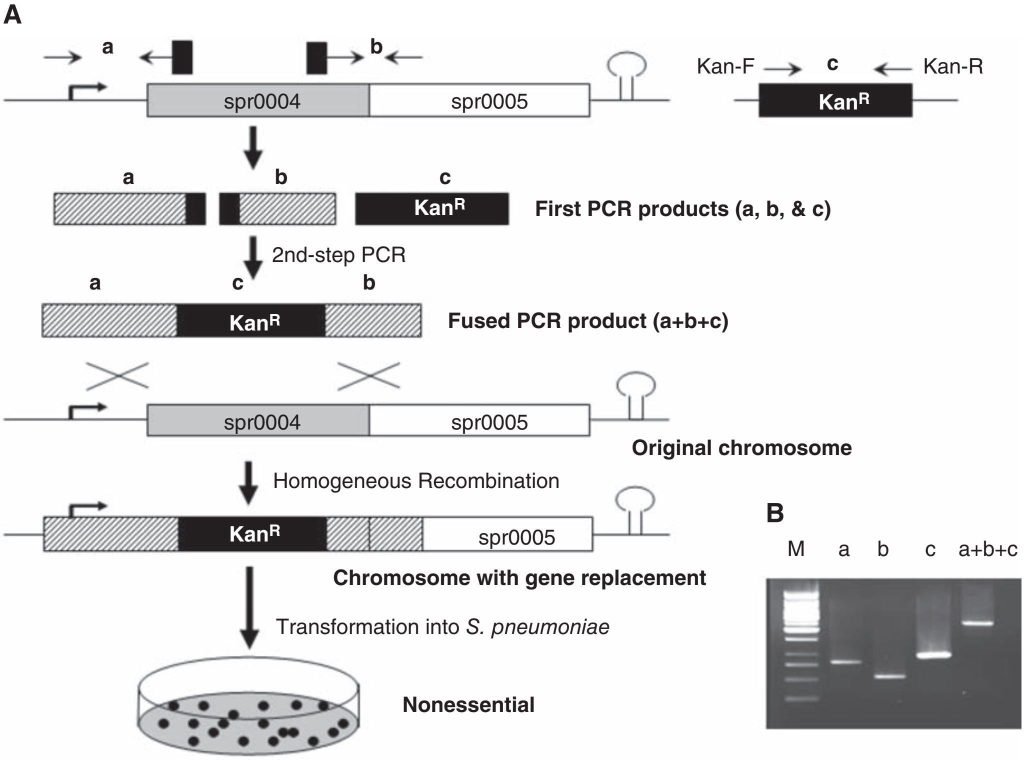

A
a
b
spr0004
spr0005
Kan-F
c
Kan-R
KanR
a
b
c
KanR
First PCR products (a, b, & c)
2nd-step PCR
a
c
b
KanR
Fused PCR product (a+b+c)
spr0004
spr0005
Original chromosome
Homogeneous Recombination
KanR
spr0005
Chromosome with gene replacement
Transformation into S. pneumoniae
Nonessential
B
M
a
b
c
a+b+c

Fig. 3. Gene knock-out of spr0004 to confirm the removal of a potential polar effect. In the first PCR reaction, the KanR gene with transcriptional termination signal removed and up- and downstream regions of spr0004 were amplified. As a result, the upstream (**a**, 813 bp) and downstream (**b**, 558 bp) regions and the KanR gene (**c**, 904 bp) were obtained. The ~2.3-kb-long fused PCR product (**a** + **b** + **c**, 2274 bp) was consequently produced by the second PCR. Upon transformation, hundreds of colonies were obtained on THYE agar plates containing kanamycin, indicating that spr0004 is nonessential.

for recombination. Fourth, this method can minimize potential polar effect and is applicable for both monocistronic and polycistronic genes.

3.7. Bacterial Essential Genes as Potential Targets for Novel Antimicrobial Agents

Identification of essential genes in bacterial pathogens can be applied to the development of new antimicrobial agents because common essential genes in diverse bacterial species could constitute novel targets for broad-spectrum antimicrobial agents (*15*). Because of the explosion in the number of available complete bacterial genome sequences, microbial genomics can be applied to evaluate the suitability of potential targets for new antimicrobial drugs, based on the criteria of "essentiality" or "selectivity" (*16*). Several studies based on genomics-driven, target-focused approaches have provided a valuable inventory of essential genes that can be used to select and validate antimicrobial agents (*3, 5, 9–11, 14, 17*). For example, peptide deformylase (PDF) inhibitors are the products of a genomics-driven approach to discovery of novel antimicrobial agents. Although the identification of new antimicrobial drug targets does not guarantee the development of new chemical compounds, it is an important first step.

Notes

1. One should use fresh competence medium for making *S. pneumoniae* competent cells.
2. Second PCR reaction may produce several bands such as linear fused PCR product, products of left and right blanks, and others. Thus, elution of the fused PCR product from an agarose gel may be necessary (identified by expected DNA fragment size).
3. Large amounts of linear fused PCR product might be necessary to achieve sufficient transformation rates. Because the optimal PCR conditions for the production of a gene replacement construct may differ from gene to gene, several sets of conditions might need to be tested.
4. The peptide pheromone CSP (Takara) is commonly used to increase pneumococcal transformation efficacy.
5. This kanamycin concentration has been empirically determined in a preliminary study not to give rise to background kanamycin resistance. If concentration of kanamycin in the medium is higher than 400 µg/mL, pneumococcal cells may acquire resistance to it due to reason(s) other than transformation with gene replacement constructs.
6. In order to minimize potential polar effect of mutagenesis, primers are designed so that flanking genes and intergenic regions, including potential promoters, would remain intact in the mutants. In addition, transcriptional termination signals are removed from kanamycin-resistance gene marker (Kan^R), and the cassettes are designed to integrate in the same orientation as the target genes to ensure transcription of the downstream ORFs.

Acknowledgments

This work was supported partly by Samsung Biomedical Research Institute (SBRI) and partly by the IMT-2000 project of the Ministry of Commerce, Industry, and Energy, Republic of Korea.

References

1. Jordan, I. K., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* **12**, 962–968.
2. Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., Arnaud, M., et al. (2003) Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4678–4683.
3. Arigoni, F., Talabot, F., Peitsch, M., Degerton, M. D., Meldrum, E., Allet, E., et al. (1998) A genome-based approach for the identification of essential bacterial genes. *Nat. Biotechnol.* **16**, 851–858.
4. Bruccoleri, R. E., Dougherty, T. J., and Davison, D. B. (1998) Concordance analysis of microbial genomes. *Nucleic Acids Res.* **16**, 4482–4486.
5. Thanassi, J. A., Hartman-Neumann, S. L., Dougherty, T. J., Dougherty, B. A., and Pucci, M. J. (2002) Identification of 113 conserved essential genes using a high-throughput gene disruption system in *Streptococcus pneumoniae*. *Nucleic Acids Res.* **30**, 3152–3162.
6. Pierce, B. J., Ianelli, F., and Pozzi, F. (2002) Construction of new unencapsulated (rough) strains of *Streptococcus pneumoniae. Res. Microbiol.* **153**, 243–247.

7. Trieu-Cuot, P., and Courvalin, P. (1983) Nucleotide sequence of the *Streptococcus faecalis* plasmid gene encoding the 3′5″-aminoglycoside phosphotransferase type III. *Gene* **23**, 331–341.
8. Havarstein, L., Coomaraswamy, G., and Morrison, D. A. (1995) An unmodified heptadecaptide pheromone induces competence for genetic transformation in *Streptococcus pneumoniae. Proc. Natl. Acad. Sci. U.S.A.* **92**, 11140–11144.
9. Akerley, B. J., Rubin, E. J., Novick, V. L., Amaya, K., Judson, N., and Mekalanos, J. J. (2002) A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 966–971.
10. Hutchison, C. A., Pterson, S. N., Gill, S. R., Cline, R. T., White, O. Fraser, C. M., et al. (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286**, 2165–2169.
11. Forsyth, R. A., Haselbeck, R. J., Ohlsen, K. L., Yamamoto, R. T., Xu, H., Trawick, J. D., et al. (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol. Microbiol.* **43**, 1387–1400.
12. Ji, Y., Zhang, B., Van Horn, S. F., Warren, P., Woodnutt, G., Burnham, M. K. R., and Rosenberg, M. (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science* **293**, 2266–2269.
13. Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2001) Comprehensive identification of conditionally essential genes in mycobacteria. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 12712–12717.
14. Song, J. H., Ko, K. S., Lee, J. Y., Baek, J. Y., Oh, W. S., Yoon, H. S., et al. (2005) Identification of essential genes in *Streptococcus pneumoniae* by allelic replacement mutagenesis. *Mol. Cells* **19**, 365–374.
15. Zhang, R., Ou, Z. Y., and Zhang, C. T. (2004) DEG: a database of essential genes. *Nucleic Acids Res.* **32**, D271–D272.
16. Sakharkar, K. R., Sakharkar, M. K., and Chow, V. T. K. (2004) A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa*. *In Silico Biol.* **4**, 355–360.
17. Ko, K. S., Lee, J. Y., Song, J. H., Baek, J. Y., Oh, W. S., Chun, J., and Yoon, H. S. (2006) Screening of essential gene in *Staphylococcus aureus* N315 using comparative genomics and allelic replacement mutagenesis. *J. Microbiol. Biotechnol.* **16**, 623–632.

29

Design and Application of Genome-Scale Reconstructed Metabolic Models

Isabel Rocha, Jochen Förster, and Jens Nielsen

Summary

In this chapter, the process for the reconstruction of genome-scale metabolic networks is described, and some of the main applications of such models are illustrated. The reconstruction process can be viewed as an iterative process where information obtained from several sources is combined to construct a preliminary set of reactions and constraints. This involves steps such as genome annotation; identification of the reactions from the annotated genome sequence and available literature; determination of the reaction stoichiometry; definition of compartmentation and assignment of localization; determination of the biomass composition; measurement, calculation, or fitting of energy requirements; and definition of additional constraints. The reaction and constraint sets, after debugging, may be integrated into a stoichiometric model that can be used for simulation using tools such as Flux Balance Analysis (**Section 3.8**). From the flux distributions obtained, physiologic parameters such as growth yields or minimal medium components can be calculated, and their distance from similar experimental data provides a basis from where the model may need to be improved.

Key Words: computer simulation, fluxome analysis, genome annotation, genome-scale reconstruction, metabolic engineering, metabolic flux analysis, metabolic models, metabolic networks.

1. Introduction

Advanced automation techniques in genome sequencing protocols have allowed the number of fully sequenced organisms to increase rapidly in the past few years: published sequenced genomes of both prokaryotic and eukaryotic organisms currently total nearly 400, and there are more than 1500 ongoing projects according to GOLD (Genomes OnLine Database: http://www.genomesonline.org/) as of June 30, 2006. The social and economic impact of such projects is expected to be high as many of those organisms have important industrial applications or represent important human pathogens.

One of the many potential applications of the sequenced and annotated genomes of microorganisms is the reconstruction of genome-scale mathematical models of metabolism by combining genome sequence data with biochemical knowledge.

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

Ideally, these models would comprise different levels of information, from reactions stoichiometry to reactions kinetics and regulatory information. However, although several projects already incorporate enough knowledge to allow dynamic simulation of well-known microorganisms *(1, 2)*, the current lack of kinetic and regulatory data for the majority of the sequenced organisms has been hampering this achievement. Nevertheless, in the absence of such information, it is still possible to accurately predict some capabilities of metabolic systems using steady-state analysis.

Genome-scale reconstructed metabolic models are based on the well-known stoichiometry of biochemical reactions and can be used for simulating *in silico* the phenotypic behavior of a microorganism under different environmental and genetic conditions, thus representing an important tool in metabolic engineering design. Other applications of metabolic models include the assignment of functions to unknown genes and the identification of candidate drug targets by the computation of the set of essential genes.

The first reconstructed genome-scale metabolic network to be published was that of *Haemophilus influenza* *(3)*, and since then several others have been made publicly available (**Table 1**), with many more yet unpublished metabolic reconstructions under way.

However, while the reconstruction of the metabolic network of an organism is likely to become a widespread procedure, starting with the fully sequenced and (partially) annotated genome sequence, it is currently far from being a standardized methodology. This is, to a certain extent, due to the lack of uniform computational tools for model reconstruction and manipulation, but primarily is due to the difficulties associated with the extraction of information other than what is available from an annotated genome sequence. Thus, the reconstruction of a metabolic network is laborious and requires a substantial manual evaluation of the stoichiometry of different reactions in the network: whereas it typically takes 10% of the reconstruction time to collect 90% of all reactions from the annotated genome sequence, the remaining 90% of the time is often spent collecting the remaining 10% of data from literature.

Rather than making an exhaustive review of several metabolic models that have been published or an overview of the reconstruction process, which can easily be found in the literature *(4–6)*, this chapter aims to provide the reader with the detailed information about the methodology required for reconstructing metabolic networks of a given microorganism.

2. Materials

2.1. Bioinformatics Databases

The process of reconstruction of the metabolic network of an organism requires a significant input from bioinformatics databases, where information regarding genome sequencing and metabolic reactions can be found. Some of the most important on-line resources that can be used are listed in **Table 2**.

2.2. Software Tools for Manipulation and Visualization of Metabolic Networks

Additionally, and although currently there is no single commonly accepted software package capable of performing all the steps of metabolic network reconstruction,

Table 1
Genome-Scale Models Available Online

Microorganism	Online Availability	References
Haemophilus influenzae	http://gcrg.ucsd.edu/organisms/hinfluenzae.html	(3)
Escherichia coli	http://gcrg.ucsd.edu/organisms/ecoli.html	(18, 43)
Helicobacter pylori	http://gcrg.ucsd.edu/organisms/hpylori.html	(44, 45)
Saccharomyces cerevisiae	http://www.cpb.dtu.dk/models/yeastmodel.html http://gcrg.ucsd.edu/organisms/yeast.html http://www.genome.org/content/vol15/issue10/images/data/1421/DC1/15_Juni_APPENDIX.xls	(13, 19, 46)
Aspergillus niger	http://blackwellpublishing.com/products/journals/suppmat/EJB/EJB3798/EJB3798sm.htm	(20)
Plasmodium falciparum	http://plasmocyc.stanford.edu	(47)
Methanococcus jannaschii	http://maine.ebi.ac.uk:1555/server.html	(48)
Mus musculus	http://pubs3.acs.org/acs/journals/supporting_information.page?in_coden=bipret&in_volume=21&in_start_page=112	(49)
Lactococcus lactis	http://www.fluxome.com/models/Lactococcus_lactis.html http://www.biomedcentral.com/1471-2105/7/296	(7, 50)
Lactobacillus plantarum	http://www.lacplantcyc.nl/ http://aem.asm.org/cgi/content/abstract/71/11/7253	(51)
Staphylococcus aureus	http://gcrg.ucsd.edu/organisms/staph.html http://www.mrw.interscience.wiley.com/suppmat/0006-3592/suppmat/	(52, 53)
Methanosarcina barkeri	http://gcrg.ucsd.edu/organisms/mbarkeri.html	(54)
Streptomyces coelicolor A3(2)	http://www.genome.org/cgi/content/full/15/6/820/DC1	(55)

Table 2
Major Online Databases and Resources That Can Be Used in the Metabolic Network Reconstruction Process

Database	Web address	Description
GOLD (Genomes Online Database)	http://www.genomesonline.org/	Monitoring of genome sequencing projects, including complete and ongoing projects around the world.
TIGR (The Institute for Genomic Research)	http://www.tigr.org/	Stores information about genomes of sequenced organisms, both conducted at TIGR or at other institutions. Allows downloading gene attribute information, from which the backbone of the metabolic network is constructed.
NCBI (National Centre for Biotechnology Information), GenBank	http://www.ncbi.nlm.nih.gov/Genbank/index.html	Contains sequence data of both microbial and higher organisms. Other NCBI databases include information on genomes of several species.
EBI (EMBL Nucleotide Database)	http://www.ebi.ac.uk/embl/index.html	Similar as GenBank, is a primary nucleotide sequence resource.
KEGG (Kyoto Encyclopedia of Genes and Genomes)	http://www.genome.ad.jp/kegg/	Database that includes all microorganisms with publicly available genome sequence. Stores both genomic and metabolic information.
BioCyc Database Collection	http://biocyc.org/	Contains several databases (like EcoCyc) that comprise genome and metabolic pathways of single organisms, and also a reference database (MetaCyc) on metabolic pathways from many organisms.

ExPASy (Expert Protein Analysis System), (Molecular Biology Server)	http://www.expasy.org/	The Swiss-Prot and TrEMBL available though ExPASy are protein sequence databases that provide organism specific annotation information. ENZYME is another functionality where enzyme-specific information can be found.
EMP (Enzymes and Metabolic Pathways) database	http://www.empproject.com/	Covers generic information on enzymes and metabolism. Also includes specific information for a limited number of organisms.
BRENDA enzyme database	http://www.brenda.uni-koeln.de/	Contains information about enzymes. It covers organism related information for most sequenced organisms.
MIPS (Munich Information Center for Protein Sequences)	http://mips.gsf.de/	Covers information about genomic structure and integration of data for several microorganisms, including *Saccharomyces cerevisiae* and other eukaryotes.
SGD (Saccharomyces Genome Database)	http://www.yeastgenome.org/.	Contains information on the molecular biology and genetics of the yeast *Saccharomyces cerevisiae.*
GeneQuiz	http://jura.ebi.ac.uk:8765/ext-genequiz/	Information that goes from protein sequence to biochemical function, using protein and DNA databases. Also gives information about the similarity to the closest homologue in the database.
TC DB (Transport Classification database)	http://tcdb.ucsd.edu/	Classification system for membrane transport proteins known as the Transporter Classification (TC) system (analogous to the Enzyme Commission system for classification of enzymes). Allows similarity searches.

including steady-state simulations and visualization of genome-scale metabolic models, there are several alternatives that allow execution of some of these tasks.

The tools described in **Table 3** are free for academic and research use and include three alternatives designed for distinct tasks. The Pathway Tools software package was developed mainly for assisting in the automated reconstruction of metabolic models and their subsequent modification and visualization, and both FluxAnalyzer and MetaFluxNet can be used for steady-state analysis of a metabolic model. These tools can accommodate large-scale models, that is, with more than 250 reactions (which is normally the case for genome-scale models), although some functionalities do not perform well for very large models. Recently, another tool for automatic reconstruction of metabolic networks from genome information was developed. The AUTOGRAPH method *(7)* combines available curated metabolic networks and gene orthology to predict a network for a given species.

There are also numerous tools adequate for conducting dynamic simulations of metabolic models, such as Cell Designer (http://www.celldesigner.org/ and Ref. *8*), the Gepasi software (http://www.gepasi.org/ and Ref. *9*), or Jdesigner and Jarnac tools (http://sbw.kgi.edu/research/sbwIntro.htm), where kinetic information for each individual reaction is included. However, these tools are not suitable for performing steady-state simulations on genome-scale models and are therefore not discussed here. Nevertheless, the computer-readable format SBML (Systems Biology Markup Language, described at http://sbml.org/index.psp and in Ref. *10*), which was initially developed for the representation of dynamic models, can be used as a common framework for stoichiometric models as well, as all the information concerning stoichiometry of reactions and additional constraints can be easily represented. This was accomplished by Segre and coauthors *(11)*, who have developed the methodology for automatic reconstruction and inspection of metabolic networks through the use of the Pathway Tools software.

Some of the tools illustrated in **Table 3** are adequate for direct visualization of metabolic interactions determined by the set of reactions included in a metabolic model. These features are very useful for the visual representation of genome-scale data sets like the transcriptome, proteome, fluxome, or metabolome. Other tools adequate for this kind of knowledge-based data visualization such as the Cytoscape software *(12)* also exist.

Additionally, and regarding the steady-state analysis of the metabolic network, it is possible to solve the linear programming problem of FBA (explained in detail in **Section 3.8**) using the Simplex algorithm available in several commercial software packages, such as MATLAB (Mathworks, Natick, MA; http://www.mathworks.com/) or LINDO (Lindo Systems Inc., Chicago, IL; http://www.lindo.com/), or free software such as the GLPK–GNU linear programming kit (http://www.gnu.org/software/glpk/glpk.html).

2.3. Tools for Predicting Enzyme Localization

Important information for metabolic network reconstruction is related with the localization of enzymes inside the cell (i.e., regarding the organelles in which those enzymes are active). Several tools are available that can deduce this information from the amino

Table 3
Comparison of Three Software Tools Developed for Construction, Inspection, Steady-State Simulation, and Visualization of Metabolic Models

	Pathway tools	Flux analyzer	MetaFluxNet
Model format	• Own model format. • Compatible with SBML.	• Own ASCII file format. • Compatible with SBML. • Software is written in MATLAB.	• Own model format. • Compatible with SBML.
Available operations	• The backbone of the model can be created automatically giving as input the results from genome annotation. From that information, the software evaluates the evidence that each one of the pathways from the MetaCyc database is present in the organism. • The model can be edited and modified.	• Analysis of topological properties (dead-end pathways, connectivity, enzyme subsets, stoichiometric matrix rank). • Metabolic flux analysis. • Flux balance analysis. • Graph theoretical pathways. • Elementary flux modes. • Minimal cut sets. • Visualization and exportation of the stoichiometric matrix.	• System analysis. • Metabolic flux analysis. • Flux balance analysis. • Easy export of the FBA or MFA problem to MATLAB and other platforms.
Visualization features	• The graphical representation of the network is automatic and has multiple levels of resolution. Depending on the resolution, one can include information from the chemical structures of substrates to the genes associated with each enzyme. • Easy to highlight individual entities in the network. • Visualization of fluxes or transcription data in the network.	• Visualization is possible by importing a graphical representation of the network. • Visualization of fluxes or transcription data in the network.	• Automatic generation of a graphical representation, but only for small networks (up to 80 metabolites). • Visualization of fluxes in the network.
References	http://bioinformatics.ai.sri.com/ptools/ *(56)*	http://www.mpi-magdeburg.mpg.de/projects/fluxanalyzer *(57)*	http://mbel.kaist.ac.kr/mfn/ *(58)*

acid sequence of the protein and from physiologic data for the corresponding organism. TargetP (http://www.cbs.dtu.dk/services/TargetP/) and PSort (http://psort.nibb.ac.jp/) are two examples of the most commonly used ones.

3. Methods

The genome-scale reconstruction of a metabolic network involves the following steps: (1) genome annotation; (2) identification of the biochemical reactions from the annotated genome sequence and available literature; (3) determination of the reaction stoichiometry including cofactor requirements; (4) definition of compartmentation and assignment of reaction localizations; (5) determination of the biomass composition; (6) measurement, calculation, or fitting of energy requirements; and (7) definition of additional constraints. These steps and several applications of the reconstructed models ***(13)*** are described in detail in the next sections.

As shown in **Figure 1**, this reconstruction process can be seen as an iterative process ***(14)***, where information obtained from several sources is combined to construct a pre-

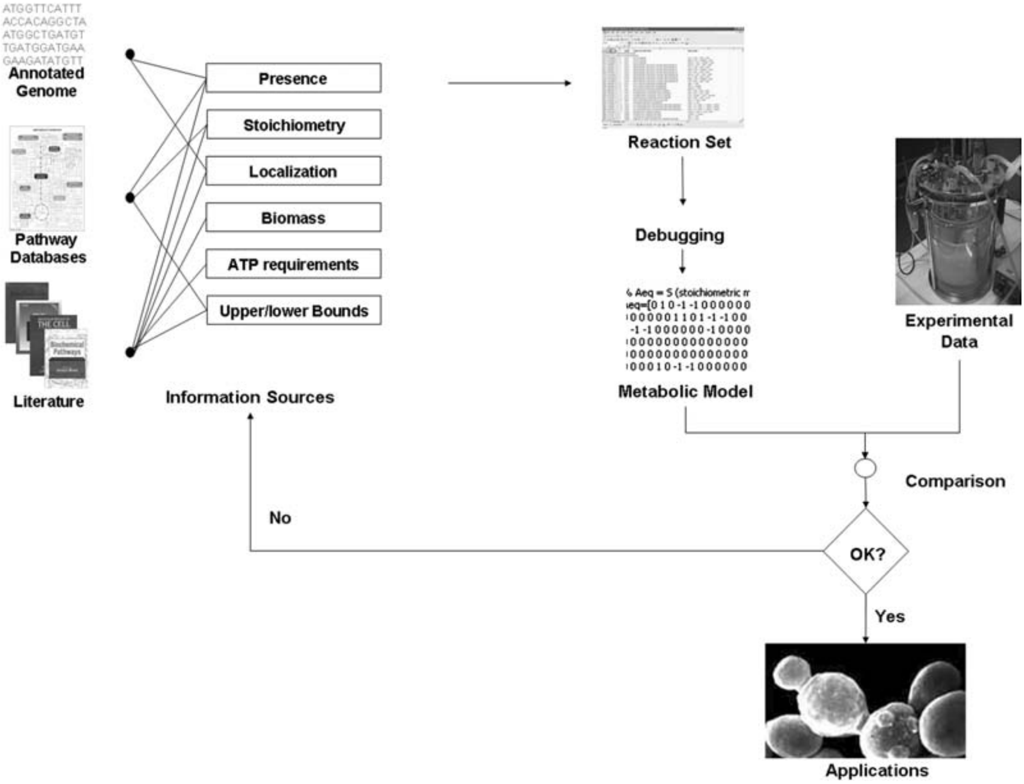


Fig. 1. Iterative process involved in the metabolic network reconstruction. The process starts with a thorough compilation of the current knowledge about the microorganism metabolism from multiple information sources. Next, the construction and debugging of the reaction set, the building of a steady-state metabolic model, and the comparison of the *in silico* simulation results with experimental data are performed. Once the experimental data are in a satisfactory agreement with the *in silico* predictions, the model can be used for further applications.

liminary set of reactions and constraints that are then analyzed to detect potential faults, for example, errors related to duplication of metabolite names, or erroneous representation of reactions catalyzed by isoenzymes and enzyme complexes.

The preliminary reaction set is further debugged and the optimized set of reactions and constraints is used to build a stoichiometric model that can be examined further by using methodologies such as Flux Balance Analysis (explained in detail in **Section 3.8**). From the flux distributions obtained using that methodology, physiologic parameters such as growth yields or minimal medium components can be calculated. Through comparison of these predictions with experimental data, the model is evaluated and is revised if necessary.

3.1. Genome Annotation

For the microorganisms with fully sequenced genomes, the process of reconstructing the metabolic network starts with a careful inspection of the data obtained from the genome annotation. More specifically, the process can be initiated by consulting a reliable public depository of genome sequence data, such as GOLD (*see* **Table 2** for additional online resources). These databases can be organism-specific or integrate information for multiple species. The EcoCyc database (http://ecocyc.org *[15]*) that stores information on the *Escherichia coli* K-12 genome and metabolism is an example of a species-specific database. TIGR or NCBI are examples of multiple-species databases.

Important data to be extracted from these sources include gene or open reading frame (ORF) names, assigned cellular functions, sequence similarities, and, for the enzyme-coding genes, the Enzyme Commission (EC) number(s) corresponding with the gene products (if available). The sequence similarity data, pertaining to the information on whether a gene is related to another gene of known function in other organisms, is usually obtained by BLAST or FASTA family algorithms (available at the NCBI and EBI Web sites; **Table 2**). The corresponding similarity scores represent the confidence level of a given gene function assignment and thus can be useful for the subsequent decision of inclusion of individual reactions in the model.

From the complete set of sequenced genes, only the genes encoding enzymes and membrane transporters are used for the reconstruction. All other annotated non-hypothetical ORFs should be scrutinized, and potential metabolic or transport reactions should be added to the reaction set. The ORFs for which no function has been yet assigned can also be analyzed with the goal of assigning a putative function, but generally one should be careful with including such ORFs into metabolic models. Genes involved in signal transduction or regulatory control of metabolic functions are currently excluded from the model-building process.

In addition, pathway databases such as KEGG, ExPASy, or the BioCyc (**Table 2**) can be utilized at this stage (via the Pathway Tools software) as they provide information about the set of reactions that have been deduced for a given microorganism from its genome sequence. However, only reactions for which an EC number has been assigned are shown, excluding both transport reactions and metabolic reactions without EC numbers. Nevertheless, these databases are important for the extraction of information about each individual reaction that has been identified as will be shown in the next sections.

3.2. Identification of Reactions

It is generally possible to speed up the initial reconstruction by using only genes that code for enzymes with EC numbers assigned. This first reaction set constitutes the backbone of the network. At the end of this process, the names of the genes assigned during genome annotation, the names of the reactions, and reactants and products for each reaction should all be included in the list of reactions.

This can be achieved, for example, by consulting the already mentioned online pathway databases, where the detailed information about each individual reaction catalyzed by a given enzyme with an assigned EC number can be obtained. Examples of such databases are BRENDA, ExPASy, KEGG, and EMP (**Table 2**).

However, the reaction set has to be further complemented with reactions catalyzed by enzymes that do not have EC numbers assigned, with transport and exchange reactions, and with reactions known to exist in a given organism, but for which no corresponding genes have been found during annotation. This can only be accomplished by thorough investigation of publications and biochemistry textbooks. The information from these sources can also serve to validate the data deduced from the genome and to discard questionable reactions with poor annotation based on low sequence similarity and those for which no evidence had been found in literature. Also, many enzymes have multiple potential reactions associated with them in public databases, but only reaction(s) specific to the organism being reconstructed should be selected based on the literature survey. Furthermore, special cases more complex than simple one-gene-to-one-enzyme-to-one-reaction relations need to be considered:

1. Many enzymes accept several different substrates, thus associating one gene with several reactions. In such cases, the name of a gene stays the same, but the names of reactions differ in the reaction set.
2. Isoenzymes are encoded by different genes, but each of them catalyzes the same reaction(s). In the reaction set, they should be considered separately, creating identical reactions associated with different genes/enzymes.
3. For reactions catalyzed by enzyme complexes, several genes are associated with one or more reactions. In most cases, every element of the complex has to be present for the reaction to take place, but there are situations where this is not true, adding additional complexity to the reaction set. Unfortunately, this kind of information is often not available and thus cannot be included in the metabolic model. Finally, even when available, the inclusion of this knowledge in the metabolic model is not straightforward, one option being the insertion of Boolean operators in a similar way as described for the incorporation of regulatory phenomena into metabolic models *(**16**)* or *(**17**)*.

Examples of metabolic models, in which most of the above-mentioned situations have been considered, are described in *(**13**)* and *(**18**)*.

3.3. Reaction Stoichiometry

Information about reaction stoichiometry can be found in the online databases such as BRENDA, ExPASy, KEGG, or EMP (**Table 2**), but only for reactions catalyzed by enzymes with assigned EC numbers. For all other reactions included in the model, stoichiometric information should be based on the literature data.

An important issue related to stoichiometry is the cofactor utilization. For many metabolic reactions, it hasn't been clarified whether enzymes require NADH or NADPH, or if both cofactors can be used. In such cases, two reactions using each of the two cofactors are usually included in the reaction list. However, an unwanted consequence of this unverified cofactor utilization is the potential appearance of net transhydrogenation reactions due to the inclusion in the model of reversible reactions that accept both cofactors, NADH and NADPH. This may lead to modeling problems, as transhydrogenation is generally unlikely to occur under physiologic conditions. One way to overcome this problem is to eliminate some of these reactions or to consider them irreversible.

3.4. Compartmentation and Localization

The distribution of metabolic reactions among different compartments inside the cell has an important impact on the performance of the metabolic network. It is important to differentiate between similar reactions that involve the same metabolite but occur in different compartments, particularly for metabolites for which there are no specific transporters and diffusion is unlikely to occur.

In prokaryotic organisms, existing compartments are largely limited to the cytosol and (in some cases) the periplasmic space, whereas in eukaryotic microorganisms, metabolic reactions can occur in many different compartments inside the cell, including mitochondrion, endoplasmic reticulum, lysosome, glyoxysome, or Golgi apparatus. For animals, it will further be necessary to differentiate between different tissues. However, localizations of enzymes (and the corresponding reactions) are often unknown, and the knowledge of transport mechanisms is currently limited to a few metabolites. At this stage, it is vital to identify important compartments, for which sufficient knowledge has been accumulated and to correctly allocate reactions associated with them. For example, in the first reconstruction of the metabolic network of *S. cerevisiae* (*19*), the compartments considered were cytosol and mitochondria, whereas in the second version of the model (*13*), the list of compartments has been extended to include peroxisomes, nucleus, endoplasmic reticulum, Golgi apparatus, and vacuole. For *Aspergillus niger* (**20**), the compartments considered were cytosol, mitochondria, and glyoxysomes. In these examples, other reactions that are known to be located in other compartments or reactions with unknown localization are usually assigned to the cytosol.

Information about reaction compartmentation can be extracted from the online databases or literature data. Alternatively, enzyme localization can be deduce from its amino acid sequence and physiologic data for the corresponding organism using software tools such as TargetP or PSort, as described in **Section 2**.

To account for the compartmentation, identical metabolites present in different compartments must be assigned different names to reflect each localization. The corresponding transport systems, if known, have to be included in the model. These can include both known and inferred transport systems. The consumption or formation of ATP and the translocation of protons, sodium, and other ions can be associated with some transport reactions, and this knowledge should be included in the model as well, whenever available.

A special case of compartmentation is the extracellular medium, where some reactions can occur, such as the interconversion between different anomers. Also, transport reactions for both metabolic reactants and products have to be included. These exchange reactions are considered inputs or outputs to the system.

3.5. Biomass Formation

In order to use a set of metabolic reaction stoichiometries for the construction of a metabolic model, it is necessary to add reactions that describe biomass formation. This can be encoded by a set of reactions that either directly denote a drain of building blocks (e.g., amino acids and nucleotides) into the biomass or, alternatively, describe a drain of macromolecules that constitute the biomass. In the latter case, reactions for the assembly of these macromolecules from various building blocks have to be specified as well. In both cases, the relative amounts of necessary macromolecules and/or building blocks are based on the biomass composition of a given species, usually available in the literature.

For p biomass constituents, the biomass formation reaction can be expressed as:

$$\sum_{k=1}^{p} c_k X_k \rightarrow \text{Biomass}, \tag{1}$$

where the c_k values are determined from the biomass composition for each metabolite or macromolecule X_k. The flux associated with this reaction represents a growth rate or a specific growth rate of an organism (**Note 1**).

The selection of one of the above-mentioned strategies depends on the availability of physiologic and biochemical data on the microorganism of interest. If such data are not available for a particular microorganism, the biomass composition has to be determined experimentally; alternatively, the biomass equation obtained for a related species can be used.

According to some authors *(21)*, the latter approach does not introduce any significant errors in model simulations, as it has been demonstrated before that a change in biomass composition merely changes the simulation results obtained with a stoichiometric model. A generic biomass equation can therefore be used, even though the biomass composition is known to change depending on physiologic conditions like the specific growth rate. However, this generalization should be taken with caution in each individual case, as for example for deletion mutants, where biomass composition changes can be significant.

3.6. Energy Requirements

Information on the growth-associated energy requirements in terms of ATP molecules needed per gram of biomass synthesized is also necessary for inclusion in the biomass equation. These requirements are related to maintenance of the membrane potential, turnover of macromolecules, and polymerization of amino acids and nucleotides. Additionally, ATP consumption for nongrowth-associated maintenance has to be considered. This can be approximated by including an irreversible reaction that converts ATP into ADP and orthophosphate. The values for these energy requirements can either be found in the literature or estimated by fitting the model results to experimental data on growth yields *(19, 22)* or other parameters.

Furthermore, the P/O ratio (relationship between ATP synthesis and oxygen consumption) for the electron transport chain needs to be defined. The operational P/O ratio can be calculated directly in detailed models, in which the electron transport and proton translocations associated with respiration are accounted for, along with the use of protons in transport and other reactions *(23)*. In less detailed models, however, it is necessary to define the operational P/O ratio *a priori.* This is often achieved by fitting the model to overall growth yields *(20)*.

3.7. Other Constraints

The main constraints to be added to a reaction set are related to the reversibility/irreversibility of the reactions (sometimes called thermodynamic constraints). This information can be found in online databases, such as BRENDA and KEGG (**Table 2**). However, this information should be used with caution as in these databases the reversibility/irreversibility is always the same regardless of which organism is considered. Often, an organism-specific literature search yields better clues on whether a reaction is reversible.

By considering fluxes through each individual reaction, these constraints can be accounted for within the order of magnitude of each flux by setting the minimum flux through a given reaction to 0 for irreversible reactions and to minus infinity for reversible reactions. If the maximal flux through a given reaction is known, it can also be added to the model as a constraint.

Also, the transport fluxes for the nutrients present in the medium can be constrained between 0 and the maximal levels. Physiologically, these constraints correspond with limited substrate availability or maximal uptake rates (**Note 2**).

The limiting substrate is usually constrained to a specific uptake rate, whereas the nonlimiting substrates are usually left unconstrained. Hence, for simulating a glucose-limiting chemostat growth on simple minimal medium (i.e., for growth of yeast), the glucose uptake is constrained to a specific uptake rate, whereas the uptake of the remaining macroelements like nitrogen, phosphate, sulfur, and oxygen remains unconstrained. However, constraining oxygen may play an important role in the simulation of aerobic chemostat cultivations in Crabtree-positive yeast; that is, yeasts that display fermentative metabolism under aerobic conditions at high growth rates and extracellular glucose concentrations *(23)*.

When metabolites are not available, the corresponding transport fluxes should be constrained to 0. The transport fluxes for metabolites that are capable of leaving the cell, such as metabolic by-products, should always be unconstrained in the outward direction.

Another important constraint is related to the reaction representing nongrowth ATP requirements (**Section 3.6**). This flux should be set to the experimentally determined or calculated rate.

3.8. Model Examination

Once the set of reactions is debugged and the stoichiometry for all the reactions is defined, one may use the set of reactions as a model to describe the function of metabolic networks in a quantitative manner. One approach may be to write dynamic mass balances for each metabolite in the network, generating a set of ordinary differential equations that may be used to simulate the dynamic behavior of metabolite concentrations. However, as mentioned in the **Introduction**, there are insufficient data on kinetic expressions and parameters, and it is therefore possible to simulate dynamic conditions for only a few pathways. A steady-state approximation is therefore generally applied, reducing the mass balances to a set of linear homogeneous equations that, for a network of M metabolites and N reactions, is expressed as:

$$\sum_{j=1}^{N} S_{ij} v_j = 0, \quad i = 1, \ldots M, \qquad (\mathbf{2})$$

where v_j corresponds with the rate of reaction j, or to the j^{th} metabolic flux, and the stoichiometric coefficient S_{ij} stands for the number of moles of metabolite i formed (or consumed) in reaction j. The stoichiometric coefficients are usually normalized so that the stoichiometric coefficient for one of the metabolites in the reaction is 1, and hereby the rate of reaction becomes equal to the rate of consumption or production of this metabolite.

In matrix notation, **equation 2** becomes:

$$S\mathbf{v} = 0. \qquad (\mathbf{3})$$

In this notation, the vector **v** stands for the flux vector, and the matrix S is the so-called stoichiometric or the $M \times N$ matrix, where each column corresponds with an individual reaction, and rows refer to the steady-state mass balances for different metabolites. In addition to internal fluxes, which are associated with chemical reactions. **v** also includes the exchange fluxes mentioned above (**Section 3.7**) that account for metabolite transport through the cell membrane (**Note 3**).

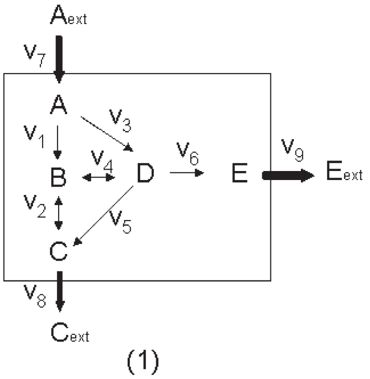
For most metabolic networks, the number of fluxes is greater than the number of mass balance constraints, resulting in an underdetermined system, for which there exists an infinite number of feasible flux distributions that satisfy the mass balance constraints, mathematically defined as the null space of S.

The constraints discussed in **Section 3.7** can be introduced as inequalities, that is,

$$\alpha_j \leq v_j \leq \beta_j, \quad j = 1, \ldots N. \qquad (\mathbf{4})$$

These constraints reduce the space of potential solutions for the system, and the intersection of the null space and the region defined by the linear inequalities has been referred to as the feasible set of solutions to a flux vector. For each given set of conditions, the cells operate at a single point within this space (corresponding with a given set of fluxes), and the corresponding flux combination must therefore be determined by other constraints not considered in the model. These may be related to gene regulatory phenomena or kinetic limitations.

As an illustration, we consider a small metabolic network of five metabolites and nine fluxes shown in **Figure 2**. Metabolite A represents one substrate available to the system, its flux being constrained to the maximum uptake rate, a, and metabolites C and E exit the system as metabolic products through an unconstrained flux. The steady-



$$
\begin{aligned}
&A \xrightarrow{v_1} B\\
&B \xleftrightarrow{v_2} 2C\\
&A \xrightarrow{v_3} D\\
&B \xleftrightarrow{v_4} D\\
&D \xrightarrow{v_5} C\\
&D \xrightarrow{v_6} E\\
&A_{ext} \xrightarrow{v_7} A\\
&C \xrightarrow{v_8} C_{ext}\\
&E \xrightarrow{v_9} E_{ext}
\end{aligned}
$$

(2)

$$
\begin{aligned}
&A: -v_1 - v_3 + v_7 = 0\\
&B: v_1 - v_2 - v_4 = 0\\
&C: 2v_2 + v_5 - v_8 = 0\\
&D: v_3 + v_4 - v_5 - v_6 = 0\\
&E: v_6 - v_9 = 0
\end{aligned}
$$

(3)

$$
\begin{aligned}
&0 \le v_1 \le +\infty\\
&-\infty \le v_2 \le +\infty\\
&0 \le v_3 \le +\infty\\
&-\infty \le v_4 \le +\infty\\
&0 \le v_5 \le +\infty\\
&0 \le v_6 \le +\infty\\
&0 \le v_7 \le a\\
&0 \le v_8 \le +\infty\\
&0 \le v_9 \le +\infty
\end{aligned}
$$

(4)

$$
\begin{bmatrix}
-1 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0\\
1 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0\\
0 & 2 & 0 & 0 & 1 & 0 & 0 & -1 & 0\\
0 & 0 & 1 & 1 & -1 & -1 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1
\end{bmatrix}
\begin{bmatrix} v_1\\ v_2\\ v_3\\ v_4\\ v_5\\ v_6\\ v_7\\ v_8\\ v_9 \end{bmatrix}
=
\begin{bmatrix} 0\\ 0\\ 0\\ 0\\ 0 \end{bmatrix}
$$

(5)

Fig. 2. Example of a metabolic network with five metabolites (**A** to **E**) and 9 fluxes (***v₁*** to ***v₉***). The reaction scheme is shown in (**1**), where the boundaries of the system are also outlined. Fluxes v_7 to v_9 represent exchange fluxes of both, metabolic substrate (**A**) and products (**C** and **E**). Reversible reactions are shown by double arrows, and irreversible reactions are indicated with a forward arrow. The stoichiometry of the network is represented in panel (**2**). Panel (**3**) shows the steady-state mass balances, and panel (**4**) illustrates the constraints around the flux values (*a* represents the maximum uptake rate for the consumption of the substrate **A**). Note that a flux value can be negative for reversible reactions with unconstrained fluxes. Panel (**5**) shows the representation of the mass balances in matrix format.

state mass balances are also shown along with the reversibility/irreversibility constraints that operate on both the internal and exchange fluxes.

Besides including regulatory or kinetic phenomena into the metabolic model, another way to reduce the feasible set to a single flux distribution is to measure several exchange fluxes in order to have a determined equation system. This approach is called metabolic flux analysis (MFA) *(24)*, which can also be combined with additional information obtained by the measurement of labelling patterns of certain internal fluxes, often referred to as metabolic network analysis (MNA) *(25)*.

Nevertheless, flux balance analysis (FBA) *(21, 23)* is currently the most commonly used methodology for calculating a single solution to the metabolic model. It allows the detailed examination of the model via the use of linear programming to determine the optimal flux distributions using a specified linear objective function, enabled by the linear nature of the feasible reaction set. FBA does not require the measurements of many fluxes but requires information on the biomass composition and on the energy requirements. Having designed a feasible model, one usually needs to specify a physiologically meaningful uptake rate or growth rate.

The linear programming formulation in FBA can be specified as:

$$\begin{aligned} &\textit{Maximize} && Z \\ &\textit{subject to} && S\mathbf{v} = 0 \\ & && \alpha_j \leq v_j \leq \beta_j, \quad j = 1, \cdots N \end{aligned} \tag{5}$$

where the maximization of a given linear objective function is subjected to the constraints expressed in the metabolic model.

For metabolic applications, the linear objective function (*Z*) to be maximized or minimized can correspond with different objectives ranging from a particular metabolic engineering design objective (e.g., optimization of a metabolite production) to the maximization of cellular growth.

Studies in several different organisms have demonstrated that their metabolic networks have evolved for the optimization of the specific growth rate under several carbon source–limiting conditions ***(26)***. Thus, the most commonly used objective function is the maximization of the biomass formation reaction rate, specified in **equation 1** (***Note 4***).

For *Saccharomyces cerevisiae*, *Aspergillus*, *Escherichia coli*, *Lactococcus lactis*, and *Mus musculus*, it has been shown that this assumption of optimality allows the calculation of phenotypic behavior, although it should be kept in mind that these simulations are highly dependent on the chosen growth- and nongrowth-dependent ATP requirements as well as on some conditions in the biomass equation.

The linear programming problem of FBA can be solved using some of the tools described in **Section 2.2**. The resulting flux distributions (**Note 5**) can then be used to examine the metabolic network. Key fluxes that represent physiologic parameters for the organism under study (i.e., the specific growth rate and growth yields) for different growth conditions are particular interesting to calculate, as these can be easily compared with experimental data. At this stage, the ATP maintenance parameters described in **Section 3.6** can be manipulated within the ranges described in the literature in order to decrease the discrepancy between calculated and experimental data. Another interesting application is the analysis of the pathways within the network known to be active under a given set of conditions (such as aerobic/anaerobic growth) and their subsequent comparison with the pathways for which the FBA analysis predicts nonzero fluxes. If inconsistencies are found, the model should be scrutinized, any potentially missing reactions added, and the erroneously included reactions removed from the reaction set (**Note 6**).

The analysis of the dead-end pathways ***(18)***, implicated by the occurrence of metabolites not connected with the overall metabolic network, can also serve to examine the consistency of a metabolic model. Their presence may be due to a misassignment of a gene function, or to an insufficient evidence for reactions linking these metabolites with the metabolic network, or even to a loss of some metabolic functions in a microorganism during evolution. The first two hypotheses should be checked by using the genome annotation information, and the model can then be updated to eliminate some of the dead-end pathways.

3.9. Applications of the Reconstructed Models

A genome-scale model can be useful in a variety of different applications, such as the prediction of phenotypic behavior under various genetic and environmental conditions, or the robustness analysis of a network by the measurement of the change in the maximal flux of the objective function when the optimal flux through any particular metabolic reaction is changed ***(27)***. Besides these examples, there are many other applications of metabolic models described in the literature that cannot be covered in this section. Thus, two examples were selected to illustrate some of the most promising developments in this field: gene deletion analysis that can be regarded as *in silico* metabolic engineering, and the extension of stoichiometry-based metabolic models to include regulatory phenomena in order to improve the quality of model-based predictions.

3.9.1. Gene-Deletion Analysis

The impacts of single-gene deletions on cell growth have previously been calculated using FBA under the assumption of optimal growth for all the published reconstructed metabolic models (e.g., in Refs. ***23*** and ***28***). For *Saccharomyces cerevisiae* and *Escherichia coli,* vast amounts of growth phenotype data have been collected in various databases (MIPS and SGD). This enabled multiple comparisons between the experimental and computed phenotypes (growth/nongrowth), which were determined to agree on average by about 80% ***(23, 28)***.

Hurst and coworkers ***(29)*** used the genome-scale model of *Saccharomyces cerevisiae* to show that although the number of indispensable genes is usually low for one environmental condition, which may indicate robustness of the metabolic network, for some other environmental conditions the number of indispensable genes has clearly been underestimated.

The growth optimization assumption may not be applicable for the simulations of phenotypic behavior of single-gene deletion mutants that have not yet undergone adaptive evolution. Church and co-workers ***(30)*** have suggested that the redirection of flux is minimal in a deletion mutant compared with a wild-type or reference strain (a methodology termed MOMA, or minimization of metabolic adjustment). Application of this assumption for the growth simulations of several *E. coli* single-gene deletion mutants has indeed improved the agreement (compared with FBA). Another methodology, termed ROOM (regulatory on/off minimization), which minimizes the number of flux changes with respect to the wild type ***(31)*** has also been suggested. The authors claim that ROOM outperforms MOMA in calculating the growth phenotypes and also in calculating intracellular flux distributions.

Although the above-mentioned approaches are aimed at the identification of essential gene sets and the model validation, there are yet few references to the use of *in silico* gene deletion or addition in metabolic engineering. Maranas and co-workers ***(32–34)*** have developed a bilevel optimization framework termed OptKnock that suggests gene *knock-out* strategies for biochemical overproduction of target metabolites, assuming that the metabolic flux distributions are governed by the objective of maximizing cellular growth. The OptKnock procedure was applied to the succinate, lactate, and 1,3-propanediol production in *E. coli* with the maximization of the biomass yield for a fixed rate of glucose uptake employed as the cellular objective.

A different approach is proposed with OptGene *(35)*, a method that uses an evolutionary algorithm to rapidly identify gene-deletion strategies for a desired phenotypic objective function. Compared with OptKnock, this framework enables the solution of large gene knock-out problems in relatively short computational time. Additionally, the proposed algorithm allows the optimization of nonlinear objective functions or incorporation of nonlinear constraints and provides a family of close to optimal solutions. The principles and utility of the OptGene algorithm were demonstrated by the identification of gene-deletion strategies in *S. cerevisiae* for improving the yield and substrate-specific productivity of three metabolites, namely, vanillin, glycerol, and succinate.

3.9.2. Integration of Regulatory Information into Genome-Scale Metabolic Models

Usually, the FBA and other flux-based simulations assume that all gene products in a metabolic reaction network are available to contribute to an optimal solution. However, it is known that all organisms possess a high level of regulation of gene expression and activity of the expressed gene product. The regulation at the transcriptional level is of special importance due to the rapid development of high-throughput techniques that allow the assessment of genome-wide expression patterns.

The incorporation of transcriptional regulation data can be accomplished by using Boolean logic equations as described in *(16)* or *(17)* and applied to small-scale *(36)* and genome-scale *(1)* metabolic models of *E. coli*. In this approach, a given flux is constrained to 0 or kept at its maximum value if a given condition or a regulatory protein is present. Operators such as AND, OR, and NOT can be used to associate different conditions in the same equation. This procedure can then be applied to simulate a dynamic process by considering a pseudo–steady state for the internal fluxes and a dynamic state for the substrate consumption, biomass production, and by-product excretion fluxes.

Whereas the above-mentioned approach is mainly based on the knowledge about the regulatory phenomena, a data-based approach was proposed by Akesson and colleagues *(37)*, where the fluxes corresponding with the enzyme-coding genes that are not expressed as revealed by the transcriptome analysis under a given set of conditions are constrained to 0. This approach can be extended to include the knowledge about subunits, assembly factors, and translational activators by constraining the corresponding flux to 0 whenever at least one of these genes is not sufficiently expressed.

Notes

1. The equation representing the biomass production is usually based on biomass composition expressed in millimoles of individual compounds or macromolecules per gram of biomass. This implies that the flux through this reaction must be expressed in grams of biomass per time unit (that can be viewed as the growth rate of a microorganism), as opposed to the fluxes through other reactions that are often expressed in millimoles per time unit. Another source of possible misinterpretations is the fact that all fluxes are often normalized to 1 g of biomass

in order to facilitate their comparisons with physiologic data. After this normalization, the flux through the biomass formation equation will be expressed in grams of biomass produced per gram of biomass per hour, which is equal to the so-called specific growth rate. However, the fluxes through all other reactions are expressed in millimoles per time unit per gram of biomass.

2. As mentioned in **Note 1**, in order to facilitate the comparison with physiologic data, the fluxes through metabolic reactions are usually normalized to 1 g of cell dry weight. However, it should be stressed that this normalization is arbitrary and is usually performed when the constraints for uptake fluxes are introduced. Thus, in terms of a metabolic model, it is irrelevant which units are used as long as they are the same for all fluxes (except the biomass production) and as long as there is coherence in conducting the mass balance computations described in **Section 3.8**.
3. **Equation 3** is often presented in the literature as $S\mathbf{v} = \mathbf{b}$. This notation is used when exchange fluxes are included in the vector **b**, especially when these fluxes are experimentally determined and are not among the variables of the system. The representation of the mass balances in matrix format for the example presented in **Figure 2** will become:

$$\begin{bmatrix} -1 & 0 & -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{bmatrix} = \begin{bmatrix} -v_7 \\ 0 \\ v_8 \\ 0 \\ v_9 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} \quad (\mathbf{6})$$

It is clear that this expression is mathematically equivalent to the one shown in **Figure 2**.

4. In genetically modified organisms, the assumption of optimal growth will not always hold. Growth of these microorganisms is better explained through the hypothesis that such strains undergo minimal redistribution of fluxes with respect to the wild-type strains *(**30**)*. Also, for growth on some unusual carbon sources, that assumption is also not valid. However, it has been demonstrated *(**38**)* that under growth selection pressure, the bacteria *E. coli* K-12 MG1655 tends to evolve from a suboptimal growth yield on glycerol to the yield predicted by FBA.
5. For the majority of metabolic models, there are multiple flux distributions that have the same value of the objective function and satisfy the model constraints, meaning that those linear programming problems have flexibility and excess capacity with respect to the constraints imposed. However, using linear programming, only one of the potential solutions is highlighted. All alternate optimal solutions can then be identified by using several methods: analysis of the reduced costs from the linear programming solution *(**27**)*, extreme pathways analysis *(**39**)*, elementary flux modes *(**40**)*, or mixed integer linear programming *(**41**)*. These multiple solutions can then be used to scrutinize metabolic regulation hypotheses by discriminating between the flux distribution options using labelling experiments for the determination of metabolic fluxes *in vivo*.
6. In addition to FBA, the network can also be analyzed through the use of convex analysis and the concepts of elementary flux modes or extreme pathways. Elementary flux modes *(**40**)* can be defined as the minimal set of enzymes that could operate at a steady state with the enzymes weighted by the relative flux they need to carry for the mode to function. The extreme pathways *(**42**)* can be regarded as systemically independent biochemical pathways that represent the edges of the steady-state flux cone.

Acknowledgments

I.R. is grateful for the financial support provided by FCT (Portuguese Science Foundation) under the postdoctoral grant BPD 11634/2002 and project recSysBio (POCI/BIO/60139/2004).

References

1. Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J., and Palsson, B. O. (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92–96.
2. Tomita, M. (2001) Whole-cell simulation: a grand challenge of the 21st century. *Trends Biotechnol.* **19**, 205–210.
3. Edwards, J. S., and Palsson, B. O. (1999) Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *J. Biol. Chem.* **274**, 17410–17416.
4. Covert, M. W., Schilling, C. H., Famili, I., Edwards, J. S., Goryanin, I. I., Selkov, E., and Palsson, B. O. (2001) Metabolic modeling of microbial strains *in silico*. *Trends Biochem. Sci.* **26**, 179–186.
5. Patil, K. R., Akesson, M., and Nielsen, J. (2004) Use of genome-scale microbial models for metabolic engineering. *Curr. Opin. Biotechnol.* **15**, 1–6.
6. Price, N. D., Papin, J. A., Schilling, C. H., and Palsson, B. O. (2003) Genome-scale microbial *in silico* models: the constraints-based approach. *Trends Biotechnol.* **21**, 162–169.
7. Notebaart, R. A., van Enckevort, F. H. J., Francke, C., Siezen, R. J., and Teusink, B. (2006) Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics* **7**, 296.
8. Kitano, H., Funahashi, A., Matsuoka, Y., and Oda, K. (2005) Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol.* **23**, 961–966.
9. Mendes, P. (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.* **22**, 361–363.
10. Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531.
11. Segre, D., Zucker, J., Katz, J., Lin, X., D'Haeseleer, P., Rindone, W. P., et al. (2003) From annotated genomes to metabolic flux models and kinetic parameter fitting. *OMICS* **7**, 301–316.
12. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504.
13. Duarte, N. C., Herrgard, M. J., and Palsson, B. O. (2004) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* **14**, 1298–1309.
14. Reed, J. L., and Palsson, B. O. (2003) Thirteen years of building constraint-based *in silico* models of *Escherichia coli. J. Bacteriol.* **185**, 2692–2699.
15. Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. et al. (2002) The MetaCyc Database. *Nucleic Acids Res.* **30**, 56–58.
16. Covert, M. W., Schilling, C. H., and Palsson, B. (2001) Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol.* **213**, 73–88.
17. Cox, S. J., Levanon, S. S., Bennett, G. N., and San, K. Y. (2005) Genetically constrained metabolic flux analysis. *Metab. Eng.* **7**, 445–456.

18. Reed, J. L., Vo, T. D., Schilling, C. H., and Palsson, B. O (2003). An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* **4**, R54.
19. Forster, J., Famili, I., Fu, P., Palsson, B. O., and Nielsen, J. (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* **13**, 244–253.
20. David, H., Akesson, M., and Nielsen, J. (2003) Reconstruction of the central carbon metabolism of *Aspergillus niger. Eur. J. Biochem.* **270**, 4243–4253.
21. Varma, A., and Palsson, B. O. (1993) Metabolic capabilities of *Escherichia coli*: II. optimal growth patterns. *J. Theor. Biol.* **165**, 503–522.
22. Varma, A., and Palsson, B. O. (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol.* **60**, 3724–3731.
23. Famili, I., Forster, J., Nielsen, J., and Palsson, B. O. (2003) *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 13134–13139.
24. Stephanopoulos, G., Aristidou, A., and Nielsen, J. (1998) *Metabolic Engineering*. San Diego: Academic Press.
25. Christensen, B., and Nielsen, J. (2000) Metabolic network analysis. A powerful tool in metabolic engineering. *Adv. Biochem. Eng. Biotechnol.* **66**, 209–231.
26. Edwards, J. S., Ibarra, R. U., and Palsson, B. O. (2001) *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* **19**, 125–130.
27. Edwards, J. S., and Palsson, B. O. (2000) Robustness analysis of the *Escherichia coli* metabolic network. *Biotechnol. Prog.* **16**, 927–939.
28. Edwards, J. S., and Palsson, B. O. (2000) Metabolic flux balance analysis and the *in silico* analysis of *Escherichia coli* K-12 gene deletions. *BMC Bioinformatics* **1**, 1.
29. Papp, B., Pal, C., and Hurst, L. D. (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429**, 661–664.
30. Segre, D., Vitkup, D., and Church, G. M. (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 15112–15117.
31. Shlomi, T., Berkman, O., and Ruppin, E. (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7695–7700.
32. Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003) OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84**, 647–657.
33. Burgard, A. P., and Maranas, C. D. (2001) Probing the performance limits of the *Escherichia coli* metabolic network subject to gene additions or deletions. *Biotechnol. Bioeng.* **74**, 364–375.
34. Pharkya, P., Burgard, A. P., and Maranas, C. D. (2003) Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock. *Biotechnol. Bioeng.* **84**, 887–899.
35. Patil, K. R., Rocha, I., Forster, J., and Nielsen, J. (2005) Evolutionary programming as a platform for *in silico* metabolic engineering. *BMC Bioinformatics* **6**, 308.
36. Covert, M. W., and Palsson, B. O. (2002) Transcriptional regulation in constraints-based metabolic models of *Escherichia coli. J. Biol. Chem.* **277**, 28058–28064.
37. Akesson, M., Forster, J., and Nielsen, J. (2004) Integration of gene expression data into genome-scale metabolic models. *Metab. Eng.* **6**, 285–293.

38. Ibarra, R. U., Edwards, J. S., and Palsson, B. O. (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* **420**, 186–189.
39. Price, N. D., Papin, J. A., and Palsson, B. O. (2002) Determination of redundancy and systems properties of the metabolic network of *Helicobacter pylori* using genome-scale extreme pathway analysis. *Genome Res.* **12**, 760–769.
40. Schuster, S., Fell, D. A., and Dandekar, T. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.* **18**, 326–332.
41. Lee, S., Phalakornkule, C., Domach, M. M., and Grossmann, I. E. (2000) Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Comput. Chem. Eng.* **24**, 711–716.
42. Schilling, C. H., Letscher, D., and Palsson, B. O. (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.* **203**, 229–248.
43. Edwards, J. S., and Palsson, B. O. (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5528–5533.
44. Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S., and Palsson, B. O. (2002) Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.* **184**, 4582–4593.
45. Thiele, I., Vo, T. D., Price, N. D., and Palsson, B. O. (2005) Expanded metabolic reconstruction of *Helicobacter pylori* (iIT341 GSM/GPR): an *in silico* genome-scale characterization of single- and double-deletion mutants. *J. Bacteriol.* **187**, 5818–5830.
46. Kuepfer, L., Sauer, U., and Blank, L. M. (2005) Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Res.* **15**, 1421–1430.
47. Yeh, W., Hanekamp, T., Tsoka, S., Karp, P. D., and Altman, R. B. (2004) Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery. *Genome Res.* **14**, 917–924.
48. Tsoka, S., Simon, D., and Ouzounis, C. A. (2004) Automated metabolic reconstruction for *Methanococcus jannaschii*. *Archaea* **1**, 223–229.
49. Sheikh, K., Forster, J., and Nielsen, L. K. (2005) Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of *Mus musculus*. *Biotechnol. Prog.* **21**, 112–121.
50. Oliveira, A. P., Nielsen, J., and Forster, J. (2005) Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC Microbiol.* **5**, 39.
51. Teusink, B., van Enckevort, F. H. J., Francke, C., Wiersma, A., Wegkamp, A., Smid, E. J., and Siezen, R. J. (2005) *In silico* reconstruction of the metabolic pathways of *Lactobacillus plantarum*: Comparing predictions of nutrient requirements with those from growth experiments. *Appl. Environ. Microbiol.* **71**, 7253–7262.
52. Becker, S. A., and Palsson, B. O. (2005) Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. *BMC Microbiol.* **5**, 8.
53. Heinemann, M., Kummel, A., Ruinatscha, R., and Panke, S. (2005) *In silico* genome-scale reconstruction and validation of the *Staphylococcus aureus* metabolic network. *Biotechnol. Bioeng.* **92**, 850–864.
54. Feist, A. M., Scholten, J. C. M., Palsson, B. O., Brockman, F., and Ideker, T. (2006) Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. *Mol. Syst. Biol.* **2**, 4.

55. Borodina, I., Krabben, P., and Nielsen, J. (2005) Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome Res.* **15**, 820–829.
56. Karp, P. D., Paley, S., and Romero, P. (2002) The Pathway Tools software. *Bioinformatics* **18** (Suppl 1), S225–S232.
57. Klamt, S., Stelling, J., Ginkel, M., and Gilles, E. D. (2003) FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. *Bioinformatics* **19**, 261–269.
58. Lee, S. Y., Lee, D.-Y., and Hong, S. H. (2003) MetaFluxNet, a program package for metabolic pathway construction and analysis, and its use in large-scale metabolic flux analysis of *Escherichia coli. Genome Inform.* **14**, 23–33.

30

Predicting Gene Essentiality Using Genome-Scale *in Silico* Models

Andrew R. Joyce and Bernhard Ø. Palsson

Summary

Genome-scale metabolic models of organisms can be reconstructed using annotated genome sequence information, well-curated databases, and primary research literature. The metabolic reaction stoichiometry and other physicochemical factors are incorporated into the model, thus imposing constraints that represent restrictions on phenotypic behavior. Based on this premise, the theoretical capabilities of the metabolic network can be assessed by using a mathematical technique known as flux balance analysis (FBA). This modeling framework, also known as the constraint-based reconstruction and analysis approach, differs from other modeling strategies because it does not attempt to predict exact network behavior. Instead, this approach uses known constraints to separate the states that a system can achieve from those that it cannot. In recent years, this strategy has been employed to probe the metabolic capabilities of a number of organisms, to generate and test experimental hypotheses, and to predict accurately metabolic phenotypes and evolutionary outcomes. This chapter introduces the constraint-based modeling approach and focuses on its application to computationally predicting gene essentiality.

Key Words: computational modeling; constraint-based reconstruction and analysis; flux balance analysis (FBA); gene essentiality prediction; metabolic phenotype; systems biology.

1. Introduction

The development of high-throughput experimental techniques in recent years has led to an explosion of genome-scale data sets for a variety of organisms. Considerable efforts have yielded complete genomic sequences for hundreds of organisms *(1)*, from which gene annotation provides a list of individual cellular components. Microarray technology affords researchers the ability to probe gene expression patterns of cells and tissues on a genome scale. Genome-wide location analysis, also known as ChIP-chip *(2)*, provides transcription factor binding site information for the entire cell. Furthermore, advances in the fields of fluxomics *(3)* and proteomics further add to the vast quantity of data currently available to researchers. Integration of these data sets to extract the most relevant information to formulate a comprehensive view of biological

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

systems is a major challenge currently facing the biological research community *(4)*. Achieving this task will require comprehensive models of cellular processes.

A prudent approach to gaining biological understanding from these complex data sets involves the development of mathematical modeling, simulation, and analysis techniques *(5)*. For many years, researchers have developed and analyzed models of biological systems via simulation, but these efforts often have been hampered by lack of complete or reliable data. Some examples of the modeling philosophies and approaches that have been pursued include deterministic kinetic modeling *(6, 7)*, stochastic modeling *(8, 9)*, and Boolean modeling *(10)*. Many of these approaches are implicitly limited by requiring knowledge of unknown parameters that are difficult or impossible to experimentally determine or approximate. Furthermore, the above approaches typically require substantial computational power, thus limiting the scale of the models that can be developed.

In recent years, however, great strides have been made in developing and using genome-scale metabolic models of a number of organisms using another modeling technique that is not subject to many of the aforementioned limitations. This approach, known as constraint-based reconstruction and analysis *(11–15)*, has been employed to generate genome-scale models for organisms from all three major branches of the tree of life. Although bacterial models dominate this growing collection, a model from archaea has recently appeared, and several eukaryotic models are also available (*see* **Note 1** and **Table 1** for an overview of existing constraint-based metabolic models).

Among other uses (*see* **Note 2** and Ref. *12*), these models have facilitated the computational investigation of gene essentiality. Flux balance analysis (FBA) *(16, 17)* is a powerful mathematical approach that uses optimization by linear programming to study the properties of metabolic networks under various conditions. When using FBA, the investigator chooses a property to optimize, such as biomass production in microbial models, and then calculates the optimal flux distribution across the metabolic model that leads to this result. Accordingly, this methodology allows the investigator to assess wild-type growth capabilities of the modeled organism. Furthermore, metabolic gene knockout strains can be simulated simply by removing associated reaction(s) from the model. By comparing predicted growth rates before and after introducing the simulated gene deletion, the gene's essentiality can be assessed (i.e., growth will be zero if the removed gene is essential for biomass production). Given that this type of analysis relies on computer simulation, computational results must be confirmed by generating and studying the effects of gene knockouts at the lab bench. However, by first investigating these situations at the computer workstation, or *in silico,* researchers can be directed to the most interesting and scientifically meaningful experiments to perform, thus limiting the amount of time spent conducting experiments of less scientific value.

In this chapter, we provide an introduction to the principles that underlie constraint-based modeling and FBA of biological systems. We give a brief but practical example to directly introduce the method and associated concepts. Furthermore, we discuss both the utility and potential shortcomings of these models in studying gene essentiality by reviewing results from several published studies. Finally, we briefly discuss additional interesting applications and some potential future directions for constraint-based modeling and analysis.

Table 1
Currently Available Constraint-Based Models

Organism	Total genes	Model genes	Model metabolites	Model reactions	Reference
Bacteria					
Bacillus subtilis	4,225	614	637	754	(91)
Escherichia coli	4,405	904	625	931	(68)
		720	438	627	(55)
Geobacter sulfurreducens	3,530	588	541	523	(71)
Haemophilus influenzae	1,775	296	343	488	(56)
		400	451	461	(92)
Helicobacter pylori	1,632	341	485	476	(58)
		291	340	388	(57)
Lactococcus lactis	2,310	358	422	621	(93)
Mannheimia succiniciproducens	2,463	335	352	373	(94)
Staphylococcus aureus	2,702	619	571	641	(70)
Streptomyces coelicolor	8,042	700	500	700	(72)
Archaea					
Methanosarcina barkeri	5,072	692	558	619	(59)
Eukarya					
Mus musculus	28,287	1,156	872	1,220	(76)
Saccharomyces cerevisiae	6,183	750	646	1,149	(62)
		672	636	1,038	(61)
		708	584	1,175	(73)
Human cardiac mitochondria	615*	298	230	189	(50)
Human red blood cell	NA	NA	39	32	(77)

This table summarizes model statistics for the models developed and published to date. *This number is based on the protein species identified in a proteomics study of the human cardiac mitochondria from which the components of the reconstruction were derived *(95)*. NA, not applicable.

2. Materials

1. Scientific literature and textbooks; for example, the PubMed database (www.pubmed.gov) and biochemical and organism-specific texts.
2. Online Genomic Databases and Resources (**Table 2**).
3. Software; for example, Microsoft Excel (office.microsoft.com), MATLAB (www.mathworks.com), Mathematica (www.wolfram.com), LINDO (www.lindo.com), GAMS (www.gams.com), and SimPheny (www.genomatica.com).

3. Methods

This section outlines the general procedure (**Fig. 1**) followed in constructing and using a constraint-based model in conjunction with FBA to computationally investigate gene essentiality. This model building and analysis procedure can be divided approximately into four successive steps:

Table 2
Online Data Resources

Data type	Resource	Description	URL
Genomic	Genomes OnLine Database (GOLD)	Repository of completed and ongoing genome projects	http://www.genomesonline.org
	The Institute for Genomic Research (TIGR)	Curated databases for microbial, plant, and human genome projects	http://www.tigr.org
	National Center for Biotechnology Information (NCBI)	Curated databases of DNA sequences as well as other data	http://www.ncbi.nlm.nih.gov
	The SEED	Database resource for genome annotations using the subsystem approach	http://www.theseed.org
Transcriptomic	Gene Expression Omnibus (GEO)	Microarray and SAGE-based genome-wide expression profiles	http://www.ncbi.nlm.nih.gov/geo
	Stanford Microarray Database (SMD)	Microarray-based genome-wide expression data	http://genome-www5.stanford.edu/
Proteomic	Expert Protein Analysis System (ExPASy)	Protein sequence, structure, and 2D PAGE data	http://au.expasy.org
	BRENDA	Enzyme functional data	http://www.brenda.uni-koeln.de/
	Open Proteomics Database (OPD)	Mass spectrometry–based proteomics data	http://bioinformatics.icmb.utexas.edu/OPD
Protein-DNA interaction	Biomolecular Network Database (BIND)	Published protein-DNA interactions	http://www.bind.ca/Action/
	Encyclopedia of DNA Elements (ENCODE)	Database of functional elements in human DNA	http://genome.ucsc.edu/encode/

Protein—protein interaction	Munich Information Center for Protein Sequences (MIPS)	Links to protein-protein interaction data and resources	http://mips.gsf.de/proj/ppi
	Database of Interacting Proteins (DIP)	Published protein-protein interactions	http://dip.doe-mbi.ucla.edu
Subcellular location	Yeast GFP-Fusion Localization Database	Genome-scale protein localization data for yeast	http://yeastgfp.ucsf.edu
Phenotype	A Systematic Annotation Package for Community Analysis of Genomes (ASAP)	Single-gene deletion phenotype microarray data for *E. coli*	http://www.genome.wisc.edu/tools/asap.htm.
	General Repository for Interaction Datasets (GRID)	Synthetic lethal interactions in yeast	http://biodata.mshri.on.ca/grid
Pathway	Kyoto Encyclopedia of Genes and Genomes (KEGG)	Pathway maps for many biological processes	http://www.genome.ad.jp/kegg/
	BioCarta	Interactive graphic models of molecular and cellular pathways	http://www.biocarta.com/genes/index.asp
Organism specific	EcoCyc	Encyclopedia of *E. coli* K-12 genes and metabolism	http://www.ecocyc.org
	Saccharomyces Genome Database (SGD)	Scientific database of the molecular biology and genetics of *S. cerevisiae*	http://www.yeastgenome.org
	BioCyc	A collection of 205 pathway/genome databases for individual organisms	http://www.biocyc.org

This table details some of the databases that store and distribute genome-scale data, gene ontological information, and organism-specific data. It should also be noted that this table is by no means comprehensive in its content but rather provides a reasonably broad sample of the data and resources that are readily accessible to researchers today. 2D-PAGE, two-dimensional polyacrylamide-gel electrophoresis; GFP, green fluorescent protein; SAGE, serial analysis of gene expression.

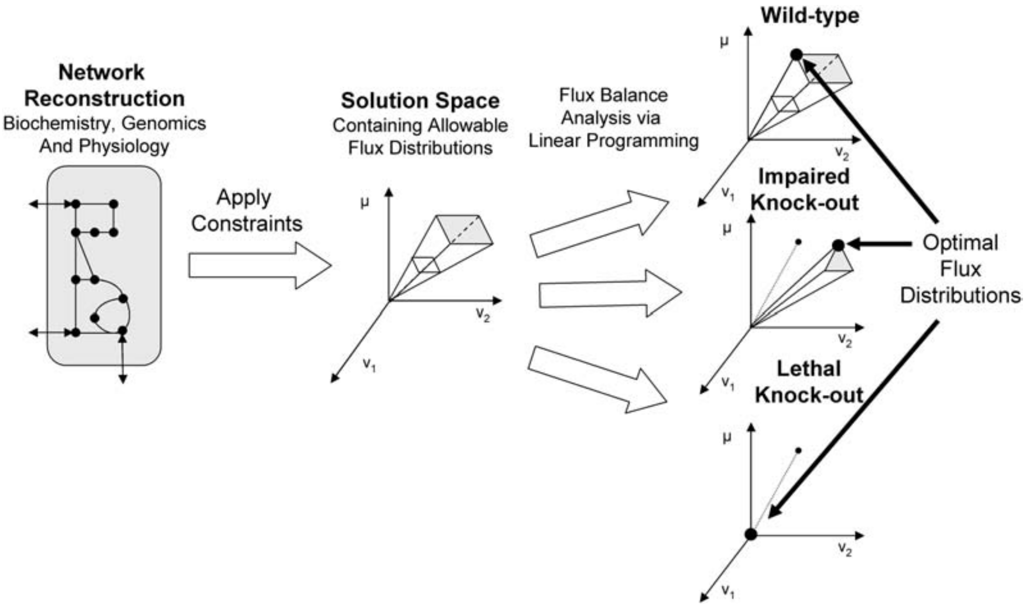

Network Reconstruction
Biochemistry, Genomics And Physiology
Apply Constraints
Solution Space
Containing Allowable Flux Distributions
Flux Balance Analysis via Linear Programming
Wild-type
Impaired Knock-out
Lethal Knock-out
Optimal Flux Distributions
μ
v_1
v_2

Fig. 1. Constraint-based modeling. Application of constraints to a reconstructed metabolic network leads to a defined solution space that specifies a cell's allowable metabolic phenotypes. Flux balance analysis (FBA) uses linear programming to find solutions in the space that maximize or minimize a given objective. In the graphical representation on the right, the optimal flux distributions that maximize μ, which represents growth/biomass production for the purposes of this chapter, are highlighted. The effects of gene knockouts on the solution space and metabolic capabilities can be assessed by simulating a gene knockout and comparing its ability to grow *in silico* relative to wild type. Impaired knockout strains are those that have a lower maximum value for the objective function than wild type, and lethal knockout strains are those that have a zero value for the objective function, indicating no growth capability when the strain harbors that particular gene deletion. As a reference, the wild-type flux distribution vector is also depicted by the dashed line on the impaired and lethal knockout plots.

1. Network reconstruction.
2. Stoichiometric (*S*) matrix compilation.
3. Identification and assignment of appropriate constraints to molecular components.
4. Assessment of gene essentiality via flux balance analysis (FBA).

In this section, each of the above components will be discussed in turn. In addition, a simple example will be provided in **Section 3.5** to illustrate directly the concepts described herein.

3.1. Network Reconstruction

The first step in constraint-based modeling, known as network reconstruction, involves generating a model that describes the system of interest. This process can be decomposed into three parts typically performed simultaneously during model construction. We detail each of these components, known individually as data collection, metabolic reaction list generation, and gene-protein-reaction (GPR) relationship determination in this section.

3.1.1. Model Component Data Collection

Perhaps the most critical component of the constraint-based modeling approach involves the collection of data that is relevant to the system of interest. Not long ago, this was among the most challenging steps as researchers had access to very limited amounts of biochemical data. However, the success of recent genome sequencing ***(18)*** and annotation ***(19, 20)*** projects and advances in high-throughput technologies as well as the development of detailed and extensive online database resources has improved matters dramatically.

After identifying the system or organism of interest, relevant data sources must be identified to begin compiling the appropriate metabolites, biochemical reactions, and associated genes to be included in the model. The three primary types of resources are the biochemical literature, high-throughput data, and integrative database resources.

3.1.1.1. Biochemical Literature

Direct biochemical information found in the primary literature usually contains the best-quality data for use in reconstructing biochemical networks. Important details, such as precise reaction stoichiometry, in addition to its reversibility, are often directly available. Given that scrutinizing each study individually is an excessively time-consuming and tedious task, biochemical textbooks and review articles should be utilized when available and the primary literature used to resolve conflicts. Furthermore, many volumes devoted to individual organisms and organelles, such as *Escherichia coli* ***(21)*** and the mitochondria ***(22)***, are increasingly becoming available and are typically excellent resources.

3.1.1.2. High-Throughput Data

Genomic and proteomic data are useful sources of information for identifying relevant metabolic network components. In recent years, the complete genome sequence for hundreds of organisms has been determined ***(18)***. Furthermore, extensive bioinformatics-based annotation efforts ***(20)*** have made great strides toward identifying all coding regions contained within the sequence. For those biochemical reactions known to occur in the organism, but whose corresponding genes are unknown, sequence alignment tools such as BLAST and FASTA ***(23)*** can be utilized to assign putative functions based on similarity to orthologous genes and proteins of known function. The subsystem approach ***(19)*** is another strategy available to researchers looking for functional gene assignments. Rather than focusing on the annotation of individual genomes, the subsystem approach calls for the annotation of cellular pathways and processes across all sequenced organisms. The associated online resource known as SEED is becoming an increasingly useful tool in constraint-based model-building efforts. It should be emphasized, however, that putative assignments are hypothetical and subject to revision upon direct biochemical characterization. As one final note on genome annotation, interesting efforts are also under way to automatically reconstruct networks based on annotated sequence information alone *(24)*. However, these automated approaches are limited in that they can only be as good as the genome annotation from which they are derived. Therefore, considerable quality-control efforts should be conducted prior to extensive use of these networks.

The proteome of a biological system defines the full complement, localization, and abundance of proteins. Although these data are generally difficult to obtain, data for some subcellular components and bacteria are available *(25, 26)*. Proteomic data are of particular importance in eukaryotic systems modeling, in which care must be taken to assign reactions to their appropriate subcellular compartment or organelle. Similarly, when modeling a system under a single condition, these data are important in identifying active components.

In addition to the primary literature, genomic and proteomic data repositories can be accessed via the Internet, as can the additional resources discussed in the next section and listed in **Table 2**.

3.1.1.3. Integrative Database Resources

In recent years, significant efforts have been devoted to developing comprehensive databases that integrate many information sources, including those data types previously described. Of particular interest are resources that have incorporated these disparate data sources into metabolic pathway maps. Kyoto Encyclopedia of Genes and Genomes (KEGG) *(27)* is perhaps the most extensive and well-known among these resource types. Pathway maps for numerous metabolic processes are available through KEGG as is information regarding orthologous genes for a variety of organisms, thus greatly enhancing the power of this resource. Additional organism-specific database resources are also available. For example, EcoCyc *(28)* incorporates gene and regulatory information as well as enzyme reaction pathways particular to *E. coli*. The Comprehensive Yeast Genome Database (CYGD) *(29)* and Saccharomyces Genome Database (SGD) *(30)* are other examples of *Saccharomyces cerevisiae*–specific comprehensive resources. Finally, the BioCyc resource *(31, 32)* contains automated annotation-derived pathway/genome databases for 250 individual organisms.

Additional important resources provide functional information for individual genes and gene products. These ontology-based tools strive to describe how gene products behave in a cellular context as they typically contain information regarding the function and localization of gene products within the cell. Perhaps the most well-known resource is Gene Ontology Consortium (GO) *(33, 34)*, which contains ontological information for a variety of organisms. In recent years, organism-specific ontologies, such as GenProtEC *(35)* for *E. coli*, have also appeared. In sum, these online resources are valuable in that they typically incorporate information regarding individual genes and proteins as well as information regarding their regulation, cellular localization, and participation in enzymatic reactions into a single integrative resource.

3.1.2. Metabolic Reaction List Generation

The next step in defining a constraint-based model requires clearly specifying the reactions to be included based on the metabolite and enzyme information collected in the previous step. A metabolic reaction can be viewed simply as substrate(s) conversion to product(s), often by enzyme-mediated catalysis. Each reaction in a metabolic network always must adhere to the fundamental laws of physics and chemistry; therefore, reactions must be balanced in terms of charge and elemental composition. For example, the depiction of the first step of glycolysis in **Figure 2A** is neither elementally nor charge balanced. However, inclusion of hydrogen in **Figure 2B** balances the reaction in both regards.

Biological boundaries also must be considered when defining reaction lists. Metabolic networks are composed of both intracellular and extracellular reactions. For example, in bacteria the reactions of glycolysis and the tricarboxylic acid cycle (TCA) take place intracellularly in the cytosol. However, glucose must be transported into the cell via an extracellular reaction in which a glucose transporter takes up extracellular glucose into the cell. An additional boundary consideration must be recognized particularly when modeling eukaryotic cells. Given that certain metabolic reactions take place in the cytosol and others take place in various organelles, reactions must be compartmentalized properly. Data that will assist in this process is now being generated in which proteins are tagged, for example, with green fluorescent protein (GFP), or recognized by antibodies and localized to subcellular compartments or organelles ***(36–38)***. Furthermore, computational tools have also been developed to predict subcellular location of proteins in eukaryotes ***(39)***.

Finally, reaction reversibility must be defined. Certain metabolic reactions can proceed in both directions. Thermodynamically, this permits reaction fluxes to take on both positive and negative values. The KEGG and BRENDA online resources (**Table 2**) are two useful resources that catalogue enzyme reversibility.

3.1.3. Determining GPR Relationships

Upon completing the reaction list, the protein or protein complexes that facilitate each metabolite substrate to product conversion must be determined. Each subunit of a protein complex must be assigned to the same reaction. Additionally, some reactions can be catalyzed by different enzymes. These so-called isozymes must all be assigned to the same appropriate reaction. Biochemical textbooks often provide the general name of the enzyme(s) responsible; however, the precise gene and associated gene product specific for the model organism of interest must be identified. The database resources detailed in **Section 3.1.1** and **Table 2** assist this process. In particular, KEGG and GO

A

$$C_6H_{12}O_6 + ATP^{3-} \xrightarrow{\text{Hexokinase}} C_6H_{11}O_6PO_3^{2-} + ADP^{2-}$$

B

$$C_6H_{12}O_6 + ATP^{3-} \xrightarrow{\text{Hexokinase}} C_6H_{11}O_6PO_3^{2-} + ADP^{2-} + H^{+}$$

Fig. 2. Charge and elementally balanced reactions. **(A)** This depiction of the hexokinase-mediated conversion of glucose to glucose-6-phosphate is neither elementally nor charge balanced. **(B)** Inclusion of hydrogen both elementally and charge balances the reaction.

provide considerable enzyme-reaction information for a variety of organisms. Furthermore, protein-protein interaction data sets, derived from yeast two-hybrid experiments (*40*), for example, may be useful resources for defining enzymatic complexes in less-defined situations. One must take care in using these data, however, given their generally high false-positive rate and questionable reproducibility (*41, 42*).

3.2. Defining the Stoichiometric Matrix

The compiled reaction list can be represented mathematically in the form of a stoichiometric (*S*) matrix. The *S* matrix is formed from the stoichiometric coefficients of the reactions that participate in a reaction network. It has $m \times n$ dimensions, where m is the number of metabolites and n is the number of reactions. Therefore, the *S* matrix is organized such that every column corresponds with a reaction, and every row corresponds with a metabolite. The *S* matrix describes how many reactions a compound participates in, and thus, how reactions are interconnected. Accordingly, each network that is reconstructed in this way effectively represents a two-dimensional annotation of the genome (*11, 43*).

Figure 3 shows how a simple two-reaction system can be represented as an *S* matrix. In this example, v_1 and v_2 denote reaction fluxes and are associated with individual proteins or protein complexes that catalyze the reactions. Element S_{ij} represents the coefficient of metabolite i in reaction j. Furthermore, notice that substrates are assigned negative coefficients and products are given positive coefficients. Also, for those reactions in which a metabolite does not participate, the corresponding element is assigned a zero value.

3.3. Identifying and Applying Appropriate Constraints

Having developed a mathematical representation of a metabolic network, the next step requires that any constraints be identified and imposed on the model. Cells are subject to a variety of constraints from environmental, physiochemical, evolutionary, and regulatory sources (*12, 14*). In and of itself, the *S* matrix defined in the previous section is a constraint in that it defines the mass and charge balance requirements for all possible metabolic reactions that are available to the cell. These stoichiometric constraints establish a geometric solution space (*see* **Fig. 1** for a graphical

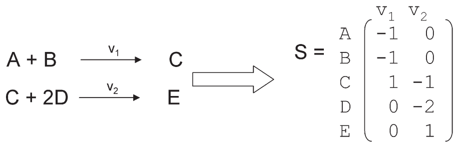


Fig. 3. Generating the stoichiometric (*S*) matrix. The reaction list on the left is mathematically represented by the *S* matrix on the right. As a convention, each row represents a metabolite, and each column represents a reaction in the network. Additionally, input or reactant metabolites have negative coefficients and outputs or products have positive coefficients. Metabolites that do not participate in a given reaction are assigned a zero value.

representation of the solution space concept) that contains all possible metabolic behaviors.

Additional constraints can be identified and imposed on the model, which has the effect of further limiting the metabolic behavior solution space. Maximum enzyme capacity ($V_{\max}$), which can be determined experimentally for some reactions, is one example and can be imposed by limiting the flux through any associated reactions to that maximum value. Furthermore, the uptake rates of certain metabolites can be determined experimentally and used to restrict metabolite uptake to the appropriate levels when mathematically analyzing the metabolic model. Additional types of constraints have also been applied, including thermodynamic limitations (*44*), internal metabolic flux determinations (*13*), and transcriptional regulation (*45–48*).

With respect to computationally assessing gene essentiality, a similar strategy to setting the maximum enzyme capacity can be utilized. By simply restricting the flux through reactions associated with the protein of interest to zero, a gene knockout can be simulated. Flux balance analysis (FBA) then can be used to examine the simulated knockout properties relative to wild type, as outlined in the next section.

3.4. Assessing Gene Essentiality via Flux Balance Analysis

Flux balance analysis (FBA) is a powerful computational method that relies on optimization by linear programming to investigate the production capabilities and systemic properties of a metabolic network. By defining an objective, such as biomass production, ATP production, or by-product secretion, FBA can be used to find an optimal flux distribution for the network model that maximizes the stated objective. This section briefly introduces some main concepts that underlie FBA, with an emphasis on how FBA can be utilized to assess gene essentiality in a metabolic network.

3.4.1. Linear Programming

The solution space defined by constraint-based models can be explored via linear optimization by utilizing linear programming (LP). The LP problem corresponding with the optimal flux distribution determination through a metabolic network can be formulated as follows:

$$\begin{aligned} &\text{Maximize} && Z = \mathbf{c}^{\mathrm{T}} v \\ &\text{Subject to} && S \cdot v = 0 \\ & && \alpha_i \le v_i \le \beta_i \quad \text{for all reactions } i. \end{aligned}$$

In the above representation, Z represents the objective function, and **c** is a vector of weights on the fluxes **v**. The weights are used to define the properties of the particular solution that is sought. The latter statements represent the flux constraints for the metabolic network. S is the matrix defined in the previous section and contains the mass and charge balanced representation of the system. Furthermore, each reaction flux v_i in the system is subject to lower and upper bound constraints, represented by α_i and β_i, respectively.

The solution to this problem yields not only a maximum value for the objective function Z, but also results in an optimal flux distribution (**v**) that allows the highest

Box 1: FBA using Matlab

Here we use Matlab to solve an FBA problem for 3 cases using the system in Figure 4. The **linprog()** function accepts six arguments and returns two values in the following form:

$$[v, Z] = \text{linprog}(c, \text{Aeq}, \text{beq}, S, b, \alpha, \beta).$$

This solves the following LP problem:

$$\begin{array}{ll} \text{Minimize} & Z = c \cdot v \\ \text{Subject to} & \text{Aeq} \cdot v \leq \text{beq} \\ & S \cdot v = b \\ & \alpha \leq v \leq \beta \end{array}$$

Since the system does not have inequality constraints other than flux vector bounds, Aeq is set equal to the identity matrix and beq to β, so that

$$\text{Aeq} \cdot v \leq \text{beq}$$

is equivalent to

$$v \leq \beta.$$

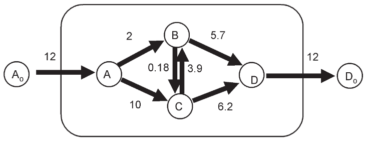
The code to solve the wild type problem (Case 1) of interest in Matlab's framework follows, using α and β as defined in the text :

```
>> S = [-1 -1 0 0 0 0 1 0;
        1 0 -1 1 -1 0 0 0;
        0 1 1 -1 0 -1 0 0;
        0 0 0 0 1 1 0 -1];
>> b = [0 0 0 0]';
>> alpha = [0 0 0 0 0 0 0 0]';
>> beta = [2 10 4 6 10 8 100 100]';
>> c = [0 0 0 0 0 0 0 1];
>> Aeq = eye(8);
>> [v,Z] = linprog(-c,Aeq,beta,S,b,alpha,beta)
Optimization terminated successfully.

v = 2.0000   10.0000    0.1822    3.9137    5.7315    6.2685
    12.0000   12.0000
Z =  -12.0000
```

Note that since Matlab defaults to solving a minimization problem we use the negative of the optimization vector.

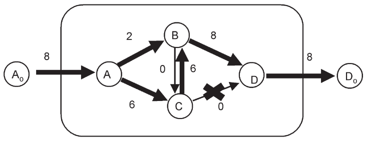
Case 1: Wild Type



Case 2 solves the same problem, but this time after knocking out reaction v5 by modifying the β vector:

```
>> beta = [2 10 4 6 10 0 100 100]';
```

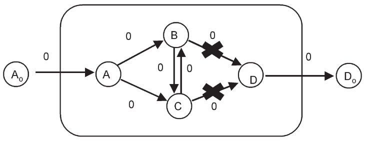
Case 2: v6 Knockout



Finally, Case 3 simulates a "lethal" deletion strain by knocking out both v5 and v6:

```
>> beta = [2 10 4 6 0 0 100 100]';
```

Case 3: v5 & v6 Double Knockout



flux through *Z*. Furthermore, computational assessment of gene essentiality is performed easily within this framework. By setting the upper and lower flux bound constraints to zero for the reaction(s) corresponding with the gene(s) of interest, a simulated gene deletion strain may be created. The examination of simulation results from before and after introducing the simulated gene deletion leads directly to gene essentiality predictions.

Problems of this type can be readily formulated and solved by commercial software packages, such as MATLAB, Mathematica, LINDO, as well as tools available through the General Algebraic Modeling System (GAMS). **Section 3.5** and **Box 1** present simple, hypothetical examples that can be solved using MATLAB. It should also be noted that these types of analyses yield a single answer; however, it is possible that multiple equivalent flux distributions that yield a maximal biomass function value exist for a given network and simulation conditions. This topic has been explored using mixed-integer linear programming (MILP) techniques with genome-scale metabolic models ***(49, 50)*** but is beyond the scope of this chapter and will not be further discussed.

3.4.2. Constraints

As previously stated, the *S* matrix constrains the system by defining all possible metabolic reactions. In mathematical terms, the stoichiometric (*S*) matrix is a linear transformation of the reaction flux vector,

$$\mathbf{v} = (v_1, v_2, \ldots, v_n)$$

to a vector of time derivatives of metabolic concentrations

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

such that

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = S \cdot \mathbf{v}.$$

Therefore, a particular flux distribution **v** represents the flux levels through each reaction in the network. Because the time constants that describe metabolic transients are fast (of the order tens of seconds or less), whereas the time constants for cell growth are comparatively slow (of the order hours to days), the behavior of cellular components can be considered as existing in a quasi-steady state *(51)*. This assumption leads to the reduction of the previous equation to:

$$S \cdot \mathbf{v} = 0.$$

By focusing only on the steady-state condition, assumptions or rough approximations regarding reaction kinetics are not needed. Furthermore, based on this premise, it is possible to determine all chemically balanced metabolic routes through the metabolic network *(52)*.

The second constraint set is imposed on the individual reaction flux values. The constraints defined by

$$\alpha_i \leq v_i \leq \beta_i \quad \text{for all reactions } i$$

specify lower and upper flux bounds for each reaction. If all model reactions are irreversible, α equals 0. Similarly, if the enzyme capacity, or $V_{\max}$, is experimentally defined, setting β to the known experimental value limits the allowable reaction flux through the enzyme within the model. In contrast, a gene knockout is simulated by setting $\beta_i = 0$ for gene i (**Section 3.5** and **Box 1**). If constraints on flux values through reaction v_i cannot be identified, then α_i and β_i are set to $-\infty$ and $+\infty$, respectively, to allow for all possible flux values. In practice, ∞ is typically represented as an arbitrarily large number that will exceed any feasible internal flux (*see* **Section 3.5** and **Box 1** for examples). Finally, if a flux is “known,” for example, from detailed experimentation, α_i and β_i can be set to the same non-zero value to explicitly define the flux value associated with reaction v_i.

A brief consideration should also be given to specifying input and output constraints on the system. When analyzing metabolic models in the context of assessing cellular growth capabilities, input constraints effectively define the environmental conditions being considered. For example, organisms have various elemental requirements that must be provided in the environment in order to support growth. Some organisms that lack certain biosynthetic processes are auxotrophic for certain biomolecules, such as amino acids, and these compounds must also be provided in the environment.

From an FBA standpoint, these issues mean that input sources must be specified in the form of input flux constraints specified in **v**. For example, if one desires to simulate rich medium conditions, flux constraints are specified such that all biomolecules that represent inputs to the system—in other words, all compounds that are available extracellularly—are left unconstrained and can flow freely into the system. In contrast, when modeling minimal medium conditions, only those inputs that are required for cell growth, or biomass formation in the formalism being considered here, are allowed to flow into the system with all other input fluxes constrained to zero (*see* Ref. ***53*** for an example of a large-scale analysis of *E. coli* growth simulations performed using minimal media). It should also be noted that certain output flux constraints may need to be set appropriately in order to allow for the simulated secretion of biomolecules that may "accumulate" in the process of forming biomass. A simple example of this is allowing for lactate and acetate secretion when modeling fermentative growth of microbes.

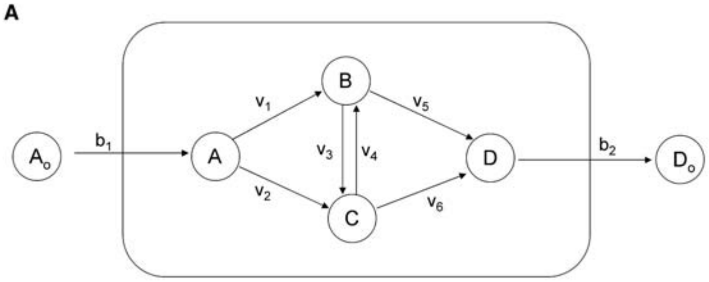
3.4.3. The Objective Function

Given that multiple possible flux distributions exist for any given network, optimization can be used to identify a particular flux distribution that maximizes or minimizes a defined objective function. Commonly used objective functions include production of ATP or production of a secreted by-product. When assessing the growth capabilities of a wild-type or simulated mutant microbe using its associated metabolic model, growth rate, as defined by the weighted consumption of metabolites needed to make biomass, is maximized. The general analysis strategy asks the question, "Is the metabolic reaction network able to support growth in the given environment, and further, is the reaction network able to support growth despite a simulated gene deletion?" Therefore, biomass generation in this modeling framework is represented as a reaction flux that drains intermediate metabolites, such as ATP, NADPH, pyruvate, and amino acids, in appropriate ratios (defined in the vector **c** of the biomass function *Z*) to support growth. As a convention, the biomass function is typically written to reflect the needs of the cell in order to make 1 g of cellular dry weight and has been experimentally determined for *E. coli* ***(54)***. In sum, with the choice of biomass as an objective function, cell growth, depicted as a non-zero value for *Z*, will only occur if all the components in the biomass function can be provided for by the network in the correct relative amounts. Accordingly, if the *in silico* knockout fails to exhibit simulated growth (i.e., $Z = 0$) (*see* **Fig. 1** for a graphical representation of this case), the associated gene is predicted to be essential.

3.5. A Simple FBA Example

In order to demonstrate the concepts previously introduced, this section presents a specific example using a simple system. **Figure 4A** shows a hypothetical four-metabolite (A, B, C, D), eight-reaction (v_1, v_2, v_3, v_4, v_5, v_6, b_1, b_2) network. By convention, each internal reaction is associated with a flux v_i, whereas reactions that span the system boundary are denoted with flux b_i. Furthermore, external metabolites A and D are denoted with subscript "o" to distinguish them from the corresponding internal metabolite. External metabolites need not be explicitly considered in the stoichiometric network representation, however.

Figure 4B outlines the reaction list associated with the system. Notice that the conversion of metabolite B to C is reversible. Rather than treating this as a single reaction,



B

```
b1:   → A
v1: A → B
v2: A → C
v3: B → C
v4: C → B
v5: B → D
v6: C → D
b2: D →
```

C

$$S=\begin{array}{c|cccccccc} & v1 & v2 & v3 & v4 & v5 & v6 & b1 & b2 \\ \hline A & -1 & -1 & 0 & 0 & 0 & 0 & 1 & 0 \\ B & 1 & 0 & -1 & 1 & -1 & 0 & 0 & 0 \\ C & 0 & 1 & 1 & -1 & 0 & -1 & 0 & 0 \\ D & 0 & 0 & 0 & 0 & 1 & 1 & 0 & -1 \end{array}$$

Fig. 4. An example system. **(A)** A four-metabolite, eight-reaction system is first decomposed into individual reactions in **(B)** and then represented mathematically in the *S* matrix depicted in **(C)**. By convention, internal reactions are denoted by v_i, and reactions that span the system boundary are denoted by b_i. External metabolites A_o and D_o need not be represented explicitly within this framework as they are outside the system under consideration.

however, for simplicity the reaction is decoupled into two separate reactions with individual corresponding fluxes.

The *S* matrix for this system is detailed in **Figure 4C**. Again, notice how this representation follows directly from the reaction list. Metabolite substrates and products are represented with negative and positive coefficients, respectively. Recall that LP problems take on the following form:

$$\begin{array}{ll} \text{Maximize} & Z = \mathbf{c}^{\mathrm{T}}\mathbf{v} \\ \text{Subject to} & S \cdot \mathbf{v} = 0 \\ & \alpha \le v_i \le \beta \quad \text{for all reactions } i. \end{array}$$

For example, if the metabolite D output is to be maximized, corresponding with maximizing the flux through b_2, the objective function is defined as follows:

$$Z = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 1) \cdot (v_1\ v_2\ v_3\ v_4\ v_5\ v_6\ b_1\ b_2)^{\mathrm{T}}$$

Furthermore, in addition to the mass and charge balance constraints imposed by the *S* matrix, lower ($\boldsymbol{\alpha}$) and upper ($\boldsymbol{\beta}$) bound vectors must be specified for the reaction vector **v**. Because all reactions in this network are irreversible, which constrains all fluxes to be positive, the lower bound vector α is set to zero:

$$\boldsymbol{\alpha} = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0)^{\mathrm{T}}$$

Upper bound values specified in vector $\boldsymbol{\beta}$ can be chosen to incorporate experimentally determined maximal enzyme capacities, also known as $V_{\max}$ values, or some arbitrarily chosen values to explore network properties. An acceptable example vector is

$$\boldsymbol{\beta} = (2\ 10\ 4\ 6\ 10\ 8\ 100\ 100)^{\mathrm{T}}.$$

The latter two upper bound values for the respective input and output fluxes are set to an arbitrarily large number in this case to reflect an effectively unlimited capacity. Accordingly, given the relatively low upper bounds on the internal fluxes, the actual values of these fluxes in the calculated optimal flux distribution will never approach these levels.

Utilizing the information compiled above, the MATLAB function **linprog()** can be used to solve for a steady-state flux distribution that maximizes for the output of metabolite D under wild-type conditions, as detailed in **Box 1**. It should be noted that the default MATLAB optimization solver is only suitable for problems of this and slightly larger magnitude. Typical biological problems that involve many more variables and constraints require more sophisticated optimization software such as the packages available through LINDO and GAMS (**Note 1**).

Having used the above information to simulate the wild-type case, the upper bound **β** vector is modified to simulate a gene deletion. For example, if we want to examine the effects of deleting the enzyme responsible for the conversion of metabolite C to D, flux v_6 is restricted to 0:

$$\boldsymbol{\beta} = (2\ 10\ 4\ 6\ 10\ \mathbf{0}\ 100\ 100)^{\mathrm{T}}.$$

Similarly, a v_5, v_6 double mutant is simulated using the following vector:

$$\boldsymbol{\beta} = (2\ 10\ 4\ 6\ \mathbf{0}\ \mathbf{0}\ 100\ 100)^{\mathrm{T}}.$$

Previous studies utilized this general strategy to simulate gene knockouts in computational investigations of gene essentiality using genome-scale bacterial models (*see*, for example, *E. coli* [*48, 55*], *H. influenzae* [*56*], *H. pylori* [*57, 58*]) as well as in the archaeal model of *M. barkeri* (*59*) and in the eukaryotic model of *S. cerevisiae* (*60–62*) (**Notes 3** and **4**).

4. Conclusion

Constraint-based modeling and its associated analyses are powerful tools that can be used to computationally predict gene essentiality with a high degree of success. This strategy aids researchers by identifying the most interesting knockouts that warrant future study, thus prioritizing experimental projects and saving considerable time. Beyond addressing the biological question associated with determining gene essentiality, this computational approach also has medical relevance. In pathogenic microbial models, each identified essential gene suggests a potential drug target that could be used to develop effective therapeutics in the future. Furthermore, progress is being made in applying this modeling framework to other aspects of the cell, such as in RNA and protein synthesis (*63*), cell signaling (*64–66*), and transcriptional regulatory networks (*67*). Because each of these network types are interrelated in terms of shared components and metabolites, these efforts are setting the stage for pushing the field a significant step forward toward generating integrated models of the entire cell (**Fig. 5**). As more genome-scale models are developed (**Note 1**), existing models enhanced (**Notes 4** and **5**), and different types of models integrated, additional applications for the constraint-based modeling approach will become apparent (**Note 2**). Consequently, the flexibility of the constraint-based modeling framework will continue to be exploited

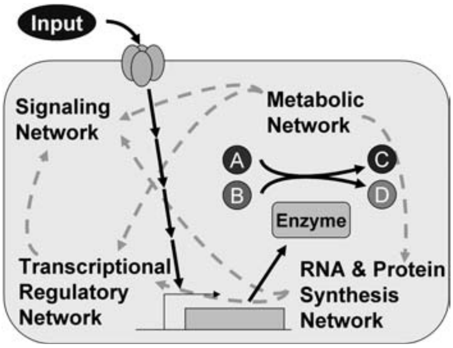

Input
Signaling Network
Metabolic Network
A
B
C
D
Enzyme
Transcriptional Regulatory Network
RNA & Protein Synthesis Network

Fig. 5. The next big challenge: model integration. This chapter has illustrated the utility of constraint-based modeling and analysis in computationally assessing gene essentiality for metabolism. The constraint-based approach has been applied to other systems as well. To date, however, these models have been developed and analyzed in isolation despite the fact that these systems are all interrelated, as shown in this conceptual figure. For example, cellular signals, or inputs, are recognized by the cell signaling network, which in turn stimulates regulatory processes. These regulatory processes mediate RNA and protein synthesis, ultimately leading to the production of enzymes that perform metabolic processes that result in cell growth or maintenance. The dashed arrows highlight the interconnectivity of these networks in the form of shared molecular components or feedback mechanisms. In principle, the constraint-based formalism can be used as a platform to capture these systems into a single picture. Accordingly, one of the next major challenges facing the field is to integrate these models of disparate cellular processes, thus pushing toward one of the field of system biology's foundational goals: to computationally represent and analyze models of entire cells and biological systems.

to aid in the prediction of gene essentiality and drive the exploration of countless other exciting biological questions.

Notes

1. This chapter presents the basic steps required to reconstruct and analyze genome-scale metabolic networks. These model systems quickly grow in size and scale, introducing computational challenges that need to be addressed. As previously noted, with large-scale models it may be necessary to use a robust computational platform designed specifically for optimization problems, such as those developed by LINDO Systems, Inc., and available through GAMS.

   Furthermore, data management becomes difficult as models scale up in size. For example, the most current *E. coli* model contains 904 genes and 931 unique biochemical reactions *(68)*. Building a genome-scale model within the framework proposed in **Section 3** is possible using ubiquitous spreadsheet software such as Excel (Microsoft, Redmond, WA), but this effort would likely be slow, unwieldy, and error-prone. In recent years, an integrative data management and analysis software platform called SimPheny (Genomatica, San Diego, CA) has been developed specifically to address the data-management and computational challenges inherent in building large-scale cellular models. This versatile platform provides network visualization, database support, and various analytical tools that greatly facilitate the construction and study of genome-scale cellular models.

Currently, more than a dozen genome-scale metabolic models have been published and are available (**Table 1**) for further research and analysis. Most of these models represent bacteria and range from the important model organism *E. coli* (***55, 68, 69***) to pathogenic microbes such as *H. pylori* (***57, 58***) and *S. aureus* (***70***). Furthermore, recently developed models of *G. sulfurreducens* (***71***) and *S. coelicolor* (***72***) may become important for their facilitation of studies that probe these organisms' respective potential bioenergetic and therapeutics-producing properties.

Representative constraint-based models have also appeared from the other two major branches of the tree of life. The recently developed metabolic reconstruction of *M. barkeri* (***59***), an interesting methanogen with bioenergetic potential, represents the first constraint-based model of an archaea that has been used to aid in the analysis of experimental data from this relatively obscure group of organisms. Furthermore, several eukaryotic models also have been developed. The metabolic models of the baker's or brewer's yeast *S. cerevisiae* (**61, 62, 73**) are second only to the *E. coli* models in terms of relative maturity and have been used in a variety of studies designed to assess network properties (for recent examples, *see* Refs. ***74*** and ***75***). Metabolic models of higher-order systems are also becoming available, such as a model of mouse (*Mus musculus* [**76**]), as well as human cardiac mitochondria (***50***) and the human red blood cell (***77***).

As more of these genome-scale models are developed, the issue of making their contents available to the broader research community is of primary concern. Given their inherent complexity, there is a need for a standardized format in which their contents can be represented in order to circumvent potential problems associated with the current typical means of distribution of models via nonstandard flat-file or spreadsheet format. In an effort to mitigate this deficiency, the Systems Biology Markup Language (SBML) (***78***), for example, has been developed to provide a uniform framework in which models can be represented, and the recently initiated MIRIAM ("minimum information requested in the annotation of biochemical models") project (***79***) and affiliated databases have appeared to provide greater transparency as to the contents and potential deficiencies of models. The adoption of these or similar standards will be important to the advancement of the field and in promoting its general utility in biological research.

2. A rapidly growing collection of analytical methods have been developed for use in conjunction with constraint-based models (reviewed in Ref. ***12***), some of which we briefly introduce in this section. Although the focus of this chapter is the use of constraint-based models to assess gene essentiality, these models can also be used to predict behavior of viable gene deletions. For example, FBA uses LP to identify the optimal metabolic state of the mutant strain. In contrast, minimization of metabolic adjustment (MOMA) uses quadratic programming (QP) to identify optimal solutions that minimize the flux distribution distance between a wild-type and simulated gene deletion strain (***86, 87***). Experimental data seem to confirm the MOMA assumption that knockout strains utilize the metabolic network similar to wild type (***86***). It remains to be determined if this is true in all situations or if the network optimizes for growth over time after gene deletion.

   A more recently developed method known as regulatory on/off minimization (ROOM) (***88***) is another constraint-based analysis technique that uses a mixed-integer linear programming (MILP) strategy to predict the metabolic state of an organism after a gene deletion by minimizing the number of flux changes that occur with respect to wild type. In other words, this algorithm aims to identify flux distributions that are qualitatively the most similar to wild type in terms of the number and types of reactions that are utilized. Whereas MOMA seems to better predict the initial metabolic adjustment that occurs after the genetic perturbation, ROOM, like FBA, better predicts the later, stabilized growth phenotype.

Constraint-based modeling also has applications in the metabolic engineering field. Identifying optimal metabolic behavior of mutant strains using a bilevel optimization framework has been employed by OptKnock *(89)*. This metabolic engineering strategy uses genome-scale metabolic models and a dual-level, nested optimization structure to predict which gene deletion(s) will lead to a desired biochemical production while retaining viable growth characteristics. This technique establishes a framework for microbial strain design and improvement *(90)* and has the potential for significant impact.

3. Many studies have used genome-scale constraint-based models to assess gene essentiality, in particular using models of *E. coli* *(48, 55)*, *H. influenzae* *(56)*, *H. pylori* *(57)*, *M. barkeri* *(59)*, and *S. cerevisiae* *(60, 62)* under various growth conditions. Each study simulated gene deletions by constraining the flux through the associated reaction(s) to zero, as described in **Section 3.4.2** and **Box 1**. Relatively few central metabolic genes are predicted to be lethal, as shown in **Table 3**. This observation likely reflects the inherent redundancy and high degree of interconnectivity that is characteristic of central metabolism. In addition, *H. influenzae* seems to be less robust than *E. coli* against single-gene deletions as a higher percentage of central metabolic genes are predicted to be essential. Furthermore, given that these networks appear generally robust against single-gene deletions, perhaps future studies should focus on lethal double mutants, known as synthetic lethal mutants, which are commonly studied in

Table 3
Computationally Predicted Gene Essentiality

Organism	No growth	Impaired growth
E. coli (*49, 55*)	*rpiAB, pgk, acnAB, gltA, icdA, tktAB, gapAC*	*atp, fba, pfkAB, tpiA, eno, gpmAB, nuo, ackAB, pta*
H. influenzae (*56*)	*eno, fba, fbp, pts, gapA, gpmA, pgi, pgk, ppc, prsA, rpiA, tktA, tpiA*	*cudABCD, atp, ndh, ackA, pta, gnd, pgl, zwf, talB, rpe*
H. pylori (*57*)	*aceB, ppa, prsA, tpi, tktA, eno*, pgi*, pgk*, gap*, pgm*, ppaA*, rpe*, rpi*, fba**	
M. barkeri (*59*)	*ackA*, pta*, cdhABCDE*, cooS*, fmdABCDEF*, fwdBDEG*, ftr*, mch*, mtd*, mer*, mtrABCDEFGH*, mtaABC*, mcrABG*, hdrABCDE*, fpoABCDFHIJKLMNO*, frhABDG*, echABCDEF*, ahaABCDEFHIK**	
S. cerevisiae (*60, 62*)	*ERG13, ACS2, ERG10, IPP1, CDS1, PSA1, TRR1, GUK1, PMI40, SAH1, SEC53, ERG26, OLE1, ERG25, ERG1, ERG11, ERG7, ERG9, ERG20, FAS1, ERG27, ERG12, ERG8, ACC1, MVD1, IDI1, FAS2, PIS1, DPM1*	*ATP16, RKI1, ILV3, ILV5, PGI1, TPI1, FBA1, PGK1*

This table summarizes some results from studies that used constraint-based metabolic models to predict gene essentiality. The “No growth” column lists the gene-deletion strains that had a simulated lethal phenotype (i.e., $Z = 0$). The “Impaired growth” column lists gene-deletion strains whose simulated phenotype was less than the wild-type strain, but not lethal (i.e., $Z_{\text{wild-type}} > Z_{\text{deletion-strain}}$).

*These genes are essential under some, but not all, tested environmental conditions.

S. cerevisiae (*80, 81*). Results from such studies are beginning to appear (*58, 61*) and may provide additional insight into gene and reaction essentiality as well as metabolic network robustness.

4. Validating model predictions is a critical component in constraint-based model analysis. Growth phenotype data, available for a number of knockout strains and organisms, can be acquired from biochemical literature (*82*) and online databases, including ASAP (*83*) for *E. coli* as well as CYGD and SGD for *S. cerevisiae*. Experimental growth phenotype data are available to assess directly the predictive power of the model for four of the five organisms listed previously and shows that correct predictions were made in ~60%, 86%, 83%, and 92% of cases for *H. pylori* (*57*), *E. coli* (*48*), *S. cerevisiae* (*62*), and *M. barkeri* (*59*), respectively. These comparisons serve two important functions: validation of the general predictive potential of the model and identification of areas that require refinement. In this sense, constraint-based models are particularly useful in experimental design by directing research to the most or least poorly understood biological components. Note 5 details how to interpret incorrect model predictions and their likely causes.
5. In the studies discussed in **Note 3** and **Note 4**, the model predictions, when compared with experimental findings, failed most often by falsely predicting growth when the gene deletion leads to a lethal phenotype *in vivo*. This trend indicates that the most common cause of false predictions is due to lack of information included in the network; for example, certain important pathways not related to metabolism in which the deleted gene participates may not be represented. In addition, the objective function may not be defined properly by failing to include the production of a compound required for growth. This latter case was shown to account for many false predictions when using a yeast metabolic model to account for strain lethality (*61*) as a few relatively minor changes to the biomass function dramatically improved the model's predictive capability. Alternatively, the gene deletion may lead to the production of a toxic by-product that ultimately kills the cell, a result for which this approach cannot account. Furthermore, certain isozymes are known to be dominant, whereas current genome-scale metabolic models typically assign equal ability to each isozyme. If this in fact is the case, the model would predict viable growth for the dominant isozyme deletion, whereas *in vivo*, the minor isozyme(s) would not sufficiently rescue the strain from the deletion of its dominant counterpart.

   An additional major error source stems from the lack of regulatory information incorporated into the previously described models. A Boolean logic approach has been used to include transcription factor–metabolic gene interactions and enhance the accuracy of constraint-based model predictions (*48*) and in genome-scale models of *E. coli* (*45*) and yeast (*84*). Regulatory information is available in the primary literature in addition to online resources such as EcoCyc and RegulonDB (*85*). Furthermore, these interactions can be derived from ChIP-chip analysis of transcription factors and corresponding gene expression microarray data (*45*).

   Incorrect predictions are less often due to false predictions of lethality. These uncommon cases often suggest the presence of previously unidentified enzyme activities, which, if added to the model, would lead to accurate predictions. They may also reflect improper biomass function definition, but in a different sense from the situation described above. For example, rather than failing to include compounds required for growth, it is also possible that certain compounds are included in the biomass function erroneously and may actually not be essential to support biological growth. In any case, inaccurate predictions often can be attributed to a paucity of information and not simply a technique failure, thus validating the general strategy of constraint-based modeling.

References

1. Wheeler, D. L., Church, D. M., Edgar, R., Federhen, S., Helmberg, W., Madden, T. L., et al. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.* **32** (Database issue), D35–40.
2. Wyrick, J. J., and Young, R. A. (2002) Deciphering gene expression regulatory networks. *Curr. Opin. Genet. Dev.* **12**, 130–136.
3. Sanford, K., Soucaille, P., Whited, G., and Chotani, G. (2002) Genomics to fluxomics and physiomics—pathway engineering. *Curr. Opin. Microbiol.* **5**, 318–322.
4. Joyce, A. R., and Palsson, B. O. (2006) The model organism as a system: integrating "omics" data sets. *Nat. Rev. Mol. Cell. Biol.* **7**, 198–210.
5. Arkin, A. P. (2001) Synthetic cell biology. *Curr. Opin. Biotechnol.* **12**, 638–644.
6. Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T. S., Matsuzaki, Y., Miyoshi, F., et al. (1999) E-CELL: software environment for whole-cell simulation. *Bioinformatics* **15**, 72–84.
7. Hoffmann, A., Levchenko, A., Scott, M. L., and Baltimore, D. (2002) The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation. *Science* **298**, 1241–1245.
8. Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002) Stochastic gene expression in a single cell. *Science* **297**, 1183–1186.
9. Arkin, A., Ross, J., and McAdams, H. H. (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* **149**, 1633–1648.
10. Sarkar, A., and Franza, B. R. (2004) A logical analysis of the process of T cell activation: different consequences depending on the state of CD28 engagement. *J. Theor. Biol.* **226**, 455–466.
11. Reed, J. L., Famili, I., Thiele, I., and Palsson, B. O. (2006) Towards multidimensional genome annotation. *Nat. Rev. Genet.* **7**, 130–141.
12. Price, N. D., Reed, J. L., and Palsson, B. O. (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2**, 886–897.
13. Edwards, J. S., Covert, M., and Palsson, B. (2002) Metabolic modelling of microbes: the flux-balance approach. *Environ. Microbiol.* **4**, 133–140.
14. Covert, M. W., Famili, I., and Palsson, B. O. (2003) Identifying constraints that govern cell behavior: a key to converting conceptual to computational models in biology? *Biotechnol. Bioeng.* **84**, 763–772.
15. Price, N. D., Papin, J. A., Schilling, C. H., and Palsson, B. O. (2003) Genome-scale microbial *in silico* models: the constraints-based approach. *Trends Biotechnol.* **21**, 162–169.
16. Varma, A., and Palsson, B. O. (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol.* **60**, 3724–3731.
17. Kauffman, K. J., Prakash, P., and Edwards, J. S. (2003) Advances in flux balance analysis. *Curr. Opin. Biotechnol.* **14**, 491–496.
18. Liolios, K., Tavernarakis, N., Hugenholtz, P., and Kyrpides, N. C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.* **34**, D332–334.
19. Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691–5702.

20. Brent, M. R. (2005) Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Res.* **15**, 1777–1786.
21. Neidhardt, F. C., and Curtiss, R. (1996) *Escherichia coli and Salmonella: cellular and molecular biology*, 2nd ed. Washington, DC: ASM Press.
22. Scheffler, I. E. (1999) *Mitochondria*. New York: Wiley-Liss.
23. Chen, Z. (2003) Assessing sequence comparison methods with the average precision criterion. *Bioinformatics* **19**, 2456–2460.
24. Karp, P. D., Paley, S., and Romero, P. (2002) The Pathway Tools software. *Bioinformatics* **18** (Suppl 1), S225–232.
25. Cash, P. (2003) Proteomics of bacterial pathogens. *Adv. Biochem. Eng. Biotechnol.* **83**, 93–115.
26. Taylor, S. W., Fahy, E., and Ghosh, S. S. (2003) Global organellar proteomics. *Trends Biotechnol.* **21**, 82–88.
27. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32** (Database issue), D277–280.
28. Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M., et al. (2002) The EcoCyc Database. *Nucleic Acids Res.* **30**, 56–58.
29. Mewes, H. W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., et al. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* **32** (Database issue), D41–44.
30. Christie, K. R., Weng, S., Balakrishnan, R., Costanzo, M. C., Dolinski, K., Dwight, S. S., et al. (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res* **32** (Database issue), D311–314.
31. Caspi, R., Foerster, H., Fulcher, C. A., Hopkinson, R., Ingraham, J., Kaipa, P., et al. (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **34**, D511–516.
32. Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahren, D., et al. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33**, 6083–6089.
33. Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., et al. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32** (Database issue), D258–261.
34. Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.* **34**, D322–326.
35. Serres, M. H., Goswami, S., and Riley, M. (2004) GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins. *Nucleic Acids Res.* **32** (Database issue), D300–302.
36. Coulton, G. (2004) Are histochemistry and cytochemistry "Omics"? *J. Mol. Histol.* **35**, 603–613.
37. Arita, M., Robert, M., and Tomita, M. (2005) All systems go: launching cell simulation fueled by integrated experimental biology data. *Curr. Opin. Biotechnol.* **16**, 344–349.
38. Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., and O'Shea, E. K. (2003) Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691.
39. Guda, C., and Subramaniam, S. (2005) pTARGET [corrected] a new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics* **21**, 3963–3969.

40. Fields, S. (2005) High-throughput two-hybrid analysis. The promise and the peril. *FEBS J.* **272**, 5391–5399.
41. Deeds, E. J., Ashenberg, O., and Shakhnovich, E. I. (2006) A simple physical model for scaling in protein-protein interaction networks. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 311–316.
42. Sprinzak, E., Sattath, S., and Margalit, H. (2003) How reliable are experimental protein-protein interaction data? *J. Mol. Biol.* **327**, 919–923.
43. Palsson, B. (2004) Two-dimensional annotation of genomes. *Nat. Biotechnol.* **22**, 1218–1219.
44. Beard, D. A., Liang, S. D., and Qian, H. (2002) Energy balance for analysis of complex metabolic networks. *Biophys. J.* **83**, 79–86.
45. Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J., and Palsson, B. O. (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92–96.
46. Covert, M. W., and Palsson, B. O. (2003) Constraints-based models: regulation of gene expression reduces the steady-state solution space. *J. Theor. Biol.* **221**, 309–325.
47. Covert, M. W., Schilling, C. H., and Palsson, B. (2001) Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol.* **213**, 73–88.
48. Covert, M. W., and Palsson, B. O. (2002) Transcriptional regulation in constraints-based metabolic models of *Escherichia coli. J. Biol. Chem.* **277**, 28058–28064.
49. Reed, J. L., and Palsson, B. O. (2004) Genome-scale in silico models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Res.* **14**, 1797–1805.
50. Vo, T. D., Greenberg, H. J., and Palsson, B. O. (2004) Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *J. Biol. Chem.* **279**, 39532–39540.
51. Palsson, B. O. (2006) *Systems Biology: Properties of Reconstructed Networks*. New York: Cambridge University Press.
52. Schilling, C. H., Letscher, D., and Palsson, B. O. (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.* **203**, 229–248.
53. Barrett, C. L., Herring, C. D., Reed, J. L., and Palsson, B. O. (2005) The global transcriptional regulatory network for metabolism in *Escherichia coli* exhibits few dominant functional states. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 19103–19108.
54. Neidhardt, F. C., Ingraham, J. L., and Schaechter, M. (1990) *Physiology of the Bacterial Cell*. Sunderland, MA: Sinauer Associates, Inc.
55. Edwards, J. S., and Palsson, B. O. (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5528–5533.
56. Schilling, C. H., and Palsson, B. O. (2000) Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *J. Theor. Biol.* **203**, 249–283.
57. Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S., and Palsson, B. O. (2002) Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.* **184**, 4582–4593.
58. Thiele, I., Vo, T. D., Price, N. D., and Palsson, B. (2005) An Expanded Metabolic Reconstruction of *Helicobacter pylori* (*i*IT341 GSM/GPR): An *in silico* genome-scale characterization of single and double deletion mutants. *J. Bacteriol.* **187**, 5818–5830.

59. Feist, A. M., Scholten, J. C. M., Palsson, B. O., Brockman, F. J., and Ideker, T. (2006) Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. *Mol. Syst. Biol.* **2**, msb4100046-E4100041-msb4100046-E4100014.
60. Forster, J., Famili, I., Palsson, B. O., and Nielsen, J. (2003) Large-scale evaluation of *in silico* gene deletions in *Saccharomyces cerevisiae*. *Omics* **7**, 193–202.
61. Kuepfer, L., Sauer, U., and Blank, L. M. (2005) Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Res.* **15**, 1421–1430.
62. Duarte, N. C., Herrgard, M. J., and Palsson, B. O. (2004) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* **14**, 1298–1309.
63. Allen, T. E., and Palsson, B. O. (2003) Sequence-based analysis of metabolic demands for protein synthesis in prokaryotes. *J. Theor. Biol.* **220**, 1–18.
64. Papin, J. A., and Palsson, B. O. (2004) Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. *J. Theor. Biol.* **227**, 283–297.
65. Papin, J. A., Hunter, T., Palsson, B. O., and Subramaniam, S. (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell. Biol.* **6**, 99–111.
66. Papin, J. A., and Palsson, B. O. (2004) The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis. *Biophys. J.* **87**, 37–46.
67. Gianchandani, E. P., Papin, J. A., Price, N. D., Joyce, A. R., and Palsson, B. O. (2006) Matrix formalism to describe functional States of transcriptional regulatory systems. *PLoS Comput. Biol.* **2**, e101.
68. Reed, J. L., Vo, T. D., Schilling, C. H., and Palsson, B. O. (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* **4**, R54.
69. Reed, J. L., and Palsson, B. O. (2003) Thirteen years of building constraint-based *in silico* models of *Escherichia coli*. *J. Bacteriol.* **185**, 2692–2699.
70. Becker, S. A., and Palsson, B. O. (2005) Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. *BMC Microbiol.* **5**, 8.
71. Mahadevan, R., Bond, D. R., Butler, J. E., Esteve-Nunez, A., Coppi, M. V., Palsson, B. O., Schilling, C. H., and Lovley, D. R. (2006) Characterization of metabolism in the Fe(III)-reducing organism *Geobacter sulfurreducens* by constraint-based modeling. *Appl. Environ. Microbiol.* **72**, 1558–1568.
72. Borodina, I., Krabben, P., and Nielsen, J. (2005) Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome Res.* **15**, 820–829.
73. Forster, J., Famili, I., Fu, P., Palsson, B. O., and Nielsen, J. (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* **13**, 244–253.
74. Almaas, E., Oltvai, Z. N., and Barabasi, A. L. (2005) The Activity Reaction Core and Plasticity of Metabolic Networks. *PLoS Comput. Biol.* **1**, e68.
75. Segre, D., DeLuna, A., Church, G. M., and Kishnoy, R. (2005) Modular epistasis in yeast metabolism. *Nat. Genet.* **37**, 77–83.
76. Sheikh, K., Forster, J., and Nielsen, L. K. (2005) Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of *Mus musculus*. *Biotechnol. Prog.* **21**, 112–121.
77. Wiback, S. J., and Palsson, B. O. (2002) Extreme pathway analysis of human red blood cell metabolism. *Biophys. J.* **83**, 808–818.

78. Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531.
79. Novere, N. L., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., et al. (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.* **23**, 1509–1515.
80. Hartwell, L. (2004) Genetics. Robust interactions. *Science* **303**, 774–775.
81. Tong, A. H., Lesage, G., Bader, G. D., Ding, H., Xu, H., Xin, X., et al. (2004) Global mapping of the yeast genetic interaction network. *Science* **303**, 808–813.
82. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., et al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, msb4100050-E4100051-msb4100050-E4100011.
83. Glasner, J. D., Liss, P., Plunkett, G. 3rd, Darling, A., Prasad, T., Rusch, M., et al. (2003) ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res.* **31**, 147–151.
84. Herrgard, M. J., Lee, B. S., Portnoy, V., and Palsson, B. O. (2006) Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. *Genome Res.* **16**, 627–635.
85. Salgado, H., Gama-Castro, S., Martinez-Antonio, A., Diaz-Peredo, E., Sanchez-Solano, F., Peralta-Gil, M., et al. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.* **32** (Database issue), D303–306.
86. Segre, D., Vitkup, D., and Church, G. M. (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 15112–15117.
87. Segre, D., Zucker, J., Katz, J., Lin, X., D'Haeseleer, P., Rindone, W. P., et al. (2003) From annotated genomes to metabolic flux models and kinetic parameter fitting. *Omics* **7**, 301–316.
88. Shlomi, T., Berkman, O., and Ruppin, E. (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7695–7700.
89. Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84**, 647–657.
90. Fong, S. S., Burgard, A. P., Herring, C. D., Knight, E. M., Blattner, F. R., Maranas, C. D., and Palsson, B. O. (2005) *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol. Bioeng.* **91**, 643–648.
91. Oh, Y.K., Palsson, B.O., Park, S.M., Schilling, C.M., and Mahadevon, R. (2007) Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *J. Biol. Chem.*, in press.
92. Edwards, J. S., and Palsson, B. O. (1999) Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *J. Biol. Chem.* **274**, 17410–17416.
93. Oliveira, A. P., Nielsen, J., and Forster, J. (2005) Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC Microbiol.* **5**, 39.
94. Hong, S. H., Kim, J. S., Lee, S. Y., In, Y. H., Choi, S. S., Rih, J. K., et al. (2004) The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens*. *Nat. Biotechnol.* **22**, 1275–1281.
95. Taylor, S. W., Fahy, E., Zhang, B., Glenn, G. M., Warnock, D. E., Wiley, S., et al. (2003) Characterization of the human heart mitochondrial proteome. *Nat. Biotechnol.* **21**, 281–286.

31

Comparative Approach to Analysis of Gene Essentiality

Andrei L. Osterman and Svetlana Y. Gerdes

All animals are equal, but some animals are more equal than others.
—G. Orwell, *Animal Farm*

A collection of chapters assembled in this volume provides an illustration of remarkable technological progress in genome-scale essentiality analysis in a variety of microbial species. In accord with other genomic techniques, one may anticipate that the volume of essentiality data will continue to grow at an accelerated pace as the technology becomes more robust and affordable. Despite substantial biological constraints associated with expansion to every new organism, the growing spectrum of techniques (illustrated in Part I of this book) has already allowed researchers to overcome the encountered problems for many diverse microbes. Moreover, the respective methods, once established in a given species, may be seamlessly expanded toward acquisition of massive data in a variety of experimental conditions. The rapid accumulation of genome-scale essentiality data, even if not matching the volume of sequencing or expression data, would soon create similar bioinformatics challenges. Indeed, such challenges are already apparent despite a relatively modest volume of currently available data.

Some of the bioinformatics aspects of gene-essentiality studies are highlighted in Part II of this book. As in other high-throughput technologies, a problem of converting experimental *observations* to reliable *assertions* of gene essentiality is the focus of the first stage of data analysis. Despite a deceivingly simple form (*essential* or *dispensable*), generation of these assignments is associated with substantial ambiguity. A specific challenge of many essentiality screens is that the actual experimental observations (of viable mutants) are obtained only for dispensable genes, whereas the essentiality is inferred from the "negative data" (inability to obtain viable mutants). Whereas this and other aspects of primary data analysis were successfully handled in a number of studies (including those presented in this book, *see* **Chapters 2** to **15**), relatively little progress was made toward efficient use of the obtained rich data to address fundamental and applied biological problems. Also in accord with other genomic techniques, we expect

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

the *comparative analysis* to play a critical role in functional interpretation of gene essentiality data. In this concluding chapter, we will briefly describe the expected implications and the first steps toward establishing this emerging approach (recently reviewed in Ref. *1*).

Which classes of biological problems may be addressed by genome-scale essentiality studies? Historically, a key motivation for such studies was a quest for anti-infective *drug targets*. This is particularly true of the first gene essentiality screens in bacterial pathogens pioneered by a number of industrial research groups *(2–7)*.[1] Nevertheless, it is quite obvious that the knowledge of which genes are essential and under which conditions would also strongly contribute to our *basic understanding of cellular networks*, pathways, mechanisms of adaptation, and so on. Mapping essential genes with unknown functions would directly impact *gene and pathway discovery* as a direct extension of historic single-gene knockout experiments, a mainstream approach of molecular genetics. Comparative gene essentiality analysis in combination with other techniques of comparative genomics impacts *evolutionary concepts,* such as a *minimal genome* abstraction that was discussed in a recent insightful essay by E. Koonin *(8)*. Essentiality analysis in metabolic modeling context (discussed in **Chapters 29** and **30**) would lead to straightforward applications in the field of *strain engineering*.

Despite the diversity of these research tasks, underlying motivations, and technical solutions, their implementation would invariably include a comparative analysis of essentiality data obtained in different growth conditions and/or in different species. The first step required to support such comparative studies is an *integration* of multiple genome-scale essentiality data sets within a genomic resource providing access to various types of functional data. Some of the first Web resources featuring essentiality data for a single model organism (e.g., PEC, or Profiling of *E. coli* Chromosome) or for multiple organisms (e.g., DEG, or Database of Essential Genes) were briefly introduced in **Chapters 26** and **27**. The first challenge of data integration is a necessary conversion of the published data obtained by different techniques and presented by a variety of notations into a chosen unified format. Although this step may seem straightforward, a substantial curation and simplification of the original data is often required as an inevitable compromise to enable a straightforward comparative analysis.

Another challenge of comparative analysis arises from the ambiguity of the term *essential gene*, which varies in meaning from *absolutely required for survival* to *substantially contributing to fitness*. It is important to realize that different techniques deliver essentiality assignments that would be closer to one or the other connotation. For example, a comparison of the two studies in *Escherichia coli* described in this book (**Chapters 6** and **11**) revealed an almost twofold difference in the number of inferred essential genes. In addition to a large common core (210 genes essential in both studies), the first study based on a random transposon mutagenesis *(9)* identified 393 essential genes that were deemed dispensable by the results of the gene-by-gene knockout study *(10)*. We believe that this difference largely reflects an aforementioned

[1] Although only a minor fraction of the obtained data has been publicly disclosed, we foresee that most of it will soon become available for the research community, as previously happened with sequenced genomes.

difference in the meaning of gene "essentiality" assessed by these two techniques. Therefore, the knowledge of the technical details and specific conditions used in every essentiality study is very important for their adequate interpretation.

All comparative essentiality studies can be roughly divided into two major categories: (1) an *intraspecies comparison* of gene essentiality in different growth conditions and (2) a *cross-species comparison.* Massive studies of the first type were performed in yeast ***(11–14)***, whereas only a few examples were published for bacterial systems ***(10, 15–18)***. The most straightforward application of such studies is the exploration of metabolic pathways and networks. The premise is that a subset of genes dispensable in *conditions A* (e.g., in the rich medium) but essential in *conditions B* (e.g., in minimal media) would implicate a group of pathways required for adaptation of the organism to the latter conditions (e.g., by compensating for the lack of certain nutrients in the minimal media).

In the two recent studies, the entire Keio collection (described in **Chapter 11**) containing 3985 *E. coli* knockout mutants viable in the rich medium was screened for the ability of individual mutants to grow on the minimal media supplemented by glucose ***(10)*** or glycerol ***(15)***. Despite certain experimental differences preventing the precise comparison of the two studies, the obtained results converged to a common set of ~100 conditionally essential genes. Further comparative analysis revealed a remarkable consistency between these experimental observations and predictions of gene essentiality obtained using a metabolic modeling approach (introduced in **Chapter 30**), providing an important cross-validation of the experimental and computational techniques. A detailed analysis of several detected inconsistencies allowed authors to refine certain aspects of the model and to generate testable hypotheses about functions and expression of individual genes and pathways ***(15)***. For example, based on the observed discrepancies between the model predictions and experimental observations related to genes involved in glycerol utilization (a sole carbon source in that study), the authors hypothesized that only one of the two possible alternative routes was functionally expressed under given conditions. This hypothesis was further supported by the difference in the expression pattern of respective genes established by focused reverse transcription–polymerase chain reaction (RT-PCR) experiments. Based on these first encouraging results, one may anticipate that an expansion of *differential essentiality screens* toward a larger variety of growth conditions and bacterial species will significantly impact our understanding of cellular networks, pathways, and individual genes.

Quite obviously, this type of comparative analysis (quantitative or qualitative) of gene essentiality data should be performed within the framework integrating high-quality genomic annotations (functional assignments of gene products), components of metabolic reconstruction (biochemical reactions and pathways), and other types of functional genomic data (e.g., gene expression). Among the first steps in this direction is an integration of the currently published genome-scale essentiality data sets in The SEED (http://theseed.uchicago.edu ***[19]***) and NMPDR genomic resources (http://www.nmpdr.org/ ***[20]***). The data are integrated in a simplified binary format that enables their seamless comparative analysis and visualization in the context of encoded subsystems and pathways. The SEED environment provides the user with access to many features and tools that support detailed exploration of genomic and functional contexts of individual genes. Some principles and examples illustrating this analysis were recently discussed in Ref. *1*.

For the purposes of functional interpretation of gene essentiality data, it is important to realize that the knowledge of conditionally essential gene products does not automatically translate to the knowledge of *essential functional roles* and respective cellular processes, such as biochemical reactions. Establishing tentative connections between these concepts is the heart of the metabolic reconstruction technology (see **Chapters 29** and **30**). The complexity of this task, even if limited in scope by metabolic pathways in a relatively well-studied model system of *E. coli*, is largely due to a substantial functional redundancy. This redundancy is manifested at various levels, including the existence of isoenzymes (homologous and, in some cases, nonhomologous proteins performing the same functional role), multifunctional proteins, and alternative pathways. It is additionally exacerbated by convoluted and often poorly understood regulatory mechanisms and by incomplete knowledge of certain areas of metabolism. On the bright side, the analysis of essentiality data within the framework of metabolic reconstruction is anticipated to be a powerful approach that would help elucidate many of these problems (as already illustrated by one of the examples above).

An intraspecies comparison of gene essentiality data has been applied also toward identification of genes specifically required for survival of bacterial pathogens in the animal model of infection (but dispensable for growth in laboratory culture). In addition to generating candidate virulence targets, the information obtained in such studies can contribute to understanding pathogenesis-related changes in metabolic and other cellular pathways. Such studies illustrated in this book (e.g., **Chapter 5**) are usually based on random transposon mutagenesis followed by populational screens. This is in contrast with conditional essentiality studies described above, where a screening of a systematic collection of knockout mutants appears to be a method of choice. Despite a substantial difference in technology and scope, many aspects of comparative analysis, such as data integration within genomic environment and projection over annotated pathways and subsystems (metabolic and nonmetabolic), are shared by both types of studies.

A *cross-species comparative analysis* of gene essentiality opens new opportunities in addressing fundamental and applied biological problems, but it also brings additional challenges. Among them, the most daunting task is establishing the accurate equivalencies between the genes in compared species. Claiming that a certain gene, or rather a gene product, is essential or dispensable in two or more distinct species, implies a knowledge of gene equivalence or *orthology* relationships between the respective genomes. In reality, this knowledge is not always available. Whereas a *homology* of gene products may be straightforwardly deduced from their sequence similarity, establishing a functional equivalence of such homologues in general requires additional evidence that may be derived from the analysis of genomic and functional context. Although many tools are available to perform such analysis on a case-by-case basis, an accurate genome-scale comparison of essentiality data would require their integration with a collection of consistent high-quality genomic annotations.

Projection of gene essentiality from model species to others is one of the central open problems of the cross-species comparative analysis. An ability to solve this fundamental problem would provide us with an ultimate validation of our understanding of cellular networks. At the same time, it would have important implications in many applied aspects of gene essentiality studies such as identification and ranking of candidate drug targets across the whole spectrum of target pathogens. The practical importance of such projection is obvious as hundreds of completely sequenced genomes of bacterial pathogens are already available compared with a handful of published genome-scale essentiality data sets. Despite the rapid progress in essentiality analysis techniques, the gap in the availability of these two types of data is only expected to grow.

It is important to emphasize that a cross-genome projection of gene essentiality may not be performed based solely on gene orthology, even if the latter could be firmly established. This hurdle originates from the existence of nonorthologous gene displacements ***(21)*** and alternative metabolic routes causing functional redundancy of pathways. For example, although all the *de novo* riboflavin biosynthesis genes (*ribABDHE*) are essential in *E. coli*, their orthologs are dispensable in *Bacillus subtilis* due to the existence of an additional riboflavin salvage pathway mediated by a specific transporter (gene *ypaA* in *B. subtilis*). *E. coli* lacks the riboflavin transport capability and has to synthesize it *de novo* despite the abundance of this vitamin in the growth media. At the same time, the gene *ribF*, encoding a bifunctional enzyme converting riboflavin to indispensable red-ox cofactors flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD), is essential in both species. (For a more detailed comparative analysis of this pathway in the context of drug target identification, see Ref. ***22***.)

The existence of various combinations of alternative routes in a variety of diverse species is characteristic of most (if not all) known metabolic pathways. In The SEED environment, they are captured as *functional variants* of *subsystems* inferred by the detailed comparative analysis of multiple sequenced genomes ***(19, 23)***. Many of these subsystems are supplemented with interactive diagrams (pathway maps) supporting an ability to display integrated essentiality data by color coding. Notably, even the simplest presentation of that kind constitutes a significant first step toward meaningful comparative analysis, interpretation, and tentative projection of essentiality between species. To continue with the example of the FMN and FAD biosynthesis subsystem,[2] it is likely that in Gram-positive pathogens *Staphylococcus aureus* and *Streptococcus pneumoniae*, a full complement of gene orthologs matching precisely the functional variant of *B. subtilis* (including the *ypaA* transporter) will also contain only one essential gene (*ribF*) encoding a bifunctional riboflavin kinase/FAD synthase. Moreover, using the same reasoning style, one may conjecture that in a related pathogen, *Streptococcus pyogenes*, which lacks the entire *de novo* biosynthesis of riboflavin but contains an ortholog of *ypaA* transporter, the latter gene is likely to be essential.

This example (additional examples are discussed in Ref. ***1***) illustrates the opportunities provided by subsystems-based comparative analysis of gene essentiality. Despite a qualitative nature and obvious limitations due to focusing on quasi-isolated

[2] For details, see SEED subsystem “FMN and FAD biosynthesis” at http://theseed.uchicago.edu/FIG/subsys.cgi?user=master:&ssa_name=FMN_and_FAD_biosynthesis&request=show_ssa.

subsystems with arbitrary boundaries, this approach provides a reasonable compromise between a dubious gene-by-gene analysis and a rigorous whole-cell modeling, which is currently feasible for only a handful of model species (*see* **Chapters 29** and **30**). Another advantage of the subsystems-based approach is that many (but not all) of its elements are expandable toward nonmetabolic subsystems that are not amenable to modeling by existing techniques.

In conclusion, it is tempting to point to some anticipated trends and directions of gene essentiality analysis technology development in the near future. Our main projection is that as the volume and quality of the available essentiality data continue to grow, a substantial effort will be allocated to the development of adequate bioinformatics resources and tools supporting efficient data analysis and interpretation. We expect a comparative approach based on integration and analysis of the essentiality data within the framework of pathways, subsystems, and whole-cell models to be in the forefront of these bioinformatics developments. We have already mentioned an anticipated expansion of conditional essentiality studies toward dozens of growth conditions in a variety of species. In addition to that, we expect the development of a more sophisticated, quantitative approach to gene essentiality analysis. To rephrase George Orwell's famous phrase, ". . . some genes are more essential than others!" It may sound like an oxymoron, but only with respect to a conventional binary view (essential vs. dispensable). However, the metrics of gene essentiality could be established to reflect the actual contribution of a given gene to organism fitness in given conditions. Depending on a particular experimental setup, this would be captured by the effect of gene inactivation on the growth rate, on the frequency of a respective mutant in a mixed population, and so on. Some of these metrics have been already used in publications and are presented in this book (*[11–13]*; *see* **Chapters 15** and **25**). A respective bioinformatics challenge would be to progress from qualitative (binary) comparative analysis toward explicit use of quantitative data. These data would be captured (and probably even acquired; *see* **Chapters 14** and **15**) in a format of microarray expression data, and they will likely be analyzed using similar techniques. Finally, the next technical breakthrough may be anticipated in the direction of *synthetic lethality* studies in at least a few model bacteria. This approach, a systematic analysis of double-knockout mutants, is already established, and it is being successfully developed in the yeast model (**Chapter 15**). The results obtained by this approach would reveal network interdependencies that are masked in single-knockout experiments. Among other implications, a synthetic lethality approach would allow us to address many problems associated with functional redundancy of genes and pathways mentioned above. Needless to say, a comparative bioinformatics analysis is expected to play a crucial role in the interpretation of these new data.

References

1. Gerdes, S., Edwards, R., Kubal, M., Fonstein, M., Stevens, R., and Osterman, A. (2006) Essential genes on metabolic maps. *Curr. Opin. Biotechnol.* **17**, 448–456.
2. Ji, Y. D., Zhang, B., Van Horn, S. F., Warren, P., Woodnutt, G., Burnham, M. K. R., and Rosenberg, M. (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science* **293**, 2266–2269.

3. Thanassi, J. A., Hartman-Neumann, S. L., Dougherty, T. J., Dougherty, B. A., and Pucci, M. J. (2002) Identification of 113 conserved essential genes using a high-throughput gene disruption system in *Streptococcus pneumoniae. Nucleic Acids Res.* **30**, 3152–3162.
4. Forsyth, R. A., Haselbeck, R. J., Ohlsen, K. L., Yamamoto, R. T., Xu, H., Trawick, J. D., et al. (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol. Microbiol.* **43**, 1387–1400.
5. Hare, R. S., Walker, S. S., Dorman, T. E., Greene, J. R., Guzman, L. M., Kenney, T. J., et al. (2001) Genetic footprinting in bacteria. *J. Bacteriol.* **183**, 1694–1706.
6. Arigoni, F., Talabot, F., Peitsch, M., Edgerton, M. D., Meldrum, E., Allet, E., et al. (1998) A genome-based approach for the identification of essential bacterial genes. *Nat. Biotechnol.* **16**, 851–856.
7. Reich, K. A., Chovan, L., and Hessler, P. (1999) Genome scanning in *Haemophilus influenzae* for identification of essential genes. *J. Bacteriol.* **181**, 4961–4968.
8. Koonin, E. V. (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* **1**, 127–136.
9. Gerdes, S., Scholle, M., Campbell, J., Balazsi, G., Ravasz, E., Daugherty, M., et al. (2003) Experimental determination and system-level analysis of essential genes in *E. coli* MG1655. *J. Bacteriol.* **185**, 5673–5684.
10. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., et al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knock-out mutants: the Keio collection. *Mol. Syst. Biol.* 10.1038/msb4100050.
11. Smith, V., Chou, K. N., Lashkari, D., Botstein, D., and Brown, P. O. (1996) Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* **274**, 2069–2074.
12. Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391.
13. Pan, X., Yuan, D. S., Xiang, D., Wang, X., Sookhai-Mahadeo, S., Bader, J. S., et al. (2004) A robust toolkit for functional profiling of the yeast genome. *Mol. Cell* **16**, 487–496.
14. Kumar, A., Seringhaus, M., Biery, M. C., Sarnovsky, R. J., Umansky, L., Piccirillo, S., et al. (2004) Large-scale mutagenesis of the yeast genome using a Tn7-derived multipurpose transposon. *Genome Res.* **14**, 1975–1986.
15. Joyce, A. R., Reed, J. L., White, A., Edwards, R., Osterman, A., Baba, T., et al. (2006) Experimental and computational assessment of conditionally essential genes in E. coli. *J. Bacteriol.* **188**, 8259–8271.
16. Winterberg, K. M., Luecke, J., Bruegl, A. S., and Reznikoff, W. S. (2005) Phenotypic screening of *Escherichia coli* K-12 Tn5 insertion libraries, using whole-genome oligonucleotide microarrays. *Appl. Environ. Microbiol.* **71**, 451–459.
17. Badarinarayana, V., Estep, P. W., Shendure, J., Edwards, J., Tavazoie, S., Lam, F., and Church, G. M.. (2001) Selection analyses of insertional mutants using subgenic-resolution arrays. *Nat. Biotechnol.* **19**, 1060–1065.
18. Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2001) Comprehensive identification of conditionally essential genes in mycobacteria. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 12712–12717.
19. Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691–702.

20. McNeil, L., Reich, C., Aziz, R., Disz, T., Edwards, R., Gerdes, S., et al. (2007) The National Microbial Pathogen Data Resource (NMPDR): A genomics platform based on subsystem annotation. *Nucleic Acids Res.* **35**, D347–353.
21. Koonin, E. V., Mushegian, A. R., and Bork, P. (1996) Non-orthologous gene displacement. *Trends Genet.* **12**, 334–336.
22. Gerdes, S., Scholle, M., D'Souza, M., Bernal, A., Baev, M., Farrell, M., et al. (2002) From genetic footprinting to antimicrobial drug targets: examples in cofactor biosynthetic pathways. *J. Bacteriol.* **184**, 4555–4572.
23. Ye, Y., Osterman, A., Overbeek, R., and Godzik, A. (2005) Automatic detection of subsystem/pathway variants in genome analysis. *Bioinformatics* **21** (Suppl 1), i478–486.

Index