

Methods in
Molecular Biology 1671

Springer Protocols



Michael Krogh Jensen
Jay D. Keasling *Editors*

Synthetic Metabolic Pathways

Methods and Protocols

 Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor

John M. Walker

School of Life and Medical Sciences

University of Hertfordshire

Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:

<http://www.springer.com/series/7651>

Synthetic Metabolic Pathways

Methods and Protocols

Edited by

Michael Krogh Jensen

*The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark,
Kongens Lyngby, Denmark*

Jay D. Keasling

Joint BioEnergy Institute, Emeryville, CA, USA

Editors

Michael Krogh Jensen
The Novo Nordisk Foundation
Center for Biosustainability
Technical University of Denmark
Kongens Lyngby, Denmark

Jay D. Keasling
Joint BioEnergy Institute
Emeryville, CA, USA

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-7294-4 ISBN 978-1-4939-7295-1 (eBook)
<https://doi.org/10.1007/978-1-4939-7295-1>

Library of Congress Control Number: 2017957981

© Springer Science+Business Media, LLC 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Humana Press imprint is published by Springer Nature
The registered company is Springer Science+Business Media, LLC
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Preface

Biobased conversion of renewable carbon sources into products related to health, food, and transportation is growing, and so is the industry evolved around it. The enormous interest in this *organism industry* relates to the accomplishments within metabolic engineering during the last 25 years, and the more recent adoption of biological research as an engineering discipline collectively referred to as synthetic biology. Both metabolic engineering and synthetic biology scale with reduced costs of DNA synthesis and sequencing, and by way of this the two scientific disciplines have bolstered both the production portfolio and the productivity of the organismal hosts used for production, also known as cell factories.

However, rational engineering of living cells and metabolic pathways for the benefit of society remains a challenge. This challenge is associated with the complex regulatory networks underlying the metabolism and evolutionary adaptation of living organisms. Additionally, given the size ranges of genomes for most production hosts, even targeting relatively narrow solution spaces in order to improve biobased production from cell factories remains a daunting challenge by the sheer number of individual cell lines and microbial strains that must be screened in order to identify the best performing cell factory. In order to tackle this challenge and enable effective and rational engineering of cell factories, metabolic engineering has adopted the Design-Build-Test-Learn paradigm from electrical engineering, largely facilitated by the methods development within synthetic biology.

By way of this volume of *Methods in Molecular Biology*, we aim to outline key steps associated with the design, building, and testing of *Synthetic metabolic pathways* for optimal cell factory performance and robustness and illustrate how data-driven learning from these steps can be used for rational cost-effective engineering of cell factories with improved performance. As such this volume is split into four parts focusing on each of these four steps of the iterative Design-Build-Test-Learn cycle related to modern cell factory engineering. In all four parts, there will be a set of chapters each starting with a description of the basic theory behind the method in question, followed by sections covering material and methods. At the end of the chapters, enumerated notes will complement the methods and materials by indicating potential pitfalls and considerations for their hands-on mitigation when using a given technique. By way of this format we expect the reader to quickly get familiarized with both classical and state-of-the-art procedures related to each of the four overarching themes of modern engineering of *Synthetic metabolic pathways*, and moreover get detailed hands-on information on the relevant procedures from thought leaders and technical experts in the field of cell factory engineering and optimization.

Reflecting the current state in the field of metabolic engineering and synthetic biology, the book is most detailed for methods related to model organisms like *Escherichia coli* and baker's yeast *Saccharomyces cerevisiae*, but the volume also includes examples of methods used for other chassis like actinomycetes and the “artificial leaf” catalyst *Sporomusa ovata*.

Together, we believe that this volume will be an indispensable part of the literature resources of laboratories focusing on the rational engineering of microbes, and it is our hope that it will find its way to laboratory bench tops worldwide.

Kongens Lyngby, Denmark
Emeryville, CA, USA

Michael Krogh Jensen
Jay D. Keasling

Contents

| | |
|---------------------------|-----------|
| <i>Preface</i> | <i>v</i> |
| <i>Contributors</i> | <i>ix</i> |

PART I PARTS AND MODELS FOR DESIGNING SYNTHETIC METABOLIC PATHWAYS

| | | |
|---|---|----|
| 1 | Parts Characterization for Tunable Protein Expression | 3 |
| | <i>Michael S. Klausen and Morten O.A. Sommer</i> | |
| 2 | Enzyme Nicotinamide Cofactor Specificity Reversal Guided by Automated Structural Analysis and Library Design | 15 |
| | <i>Jackson K.B. Cabn, Sabine Brinkmann-Chen, and Frances H. Arnold</i> | |
| 3 | Bacterial Genome Editing Strategy for Control of Transcription and Protein Stability | 27 |
| | <i>Ida Lauritsen, Virginia Martínez, Carlotta Ronda, Alex Toftgaard Nielsen, and Morten H.H. Nørholm</i> | |
| 4 | An Automated Pipeline for Engineering Many-Enzyme Pathways: Computational Sequence Design, Pathway Expression-Flux Mapping, and Scalable Pathway Optimization | 39 |
| | <i>Sean M. Halper, Daniel P. Cetnar, and Howard M. Salis</i> | |
| 5 | Computational Approaches on Stoichiometric and Kinetic Modeling for Efficient Strain Design | 63 |
| | <i>Mohammad Mazharul Islam and Rajib Saha</i> | |
| 6 | Extended Metabolic Space Modeling | 83 |
| | <i>Pablo Carbonell, Baudoin Delépine, and Jean-Loup Faulon</i> | |
| 7 | Computational Methods to Assess the Production Potential of Bio-Based Chemicals | 97 |
| | <i>Miguel A. Campodonico, Sumesh Sukumara, Adam M. Feist, and Markus J. Herrgård</i> | |

PART II BUILDING AND REFACTORIZING SYNTHETIC METABOLIC PATHWAYS

| | | |
|----|---|-----|
| 8 | Multiplex Genome Editing in <i>Escherichia coli</i> | 119 |
| | <i>Sheila Ingemann Jensen and Alex Toftgaard Nielsen</i> | |
| 9 | Designing and Implementing Algorithmic DNA Assembly Pipelines for Multi-Gene Systems | 131 |
| | <i>Szu-Yi Hsu and Michael J. Smanski</i> | |
| 10 | An Adaptive Laboratory Evolution Method to Accelerate Autotrophic Metabolism | 149 |
| | <i>Tian Zhang and Pier-Luc Tremblay</i> | |

| | | |
|--|---|-----|
| 11 | CRISPR-Cas9 Toolkit for Actinomycete Genome Editing | 163 |
| | <i>Yaojun Tong, Helene Lunde Robertsen, Kai Blin, Tilmann Weber, and Sang Yup Lee</i> | |
| 12 | Assembly and Multiplex Genome Integration of Metabolic Pathways in Yeast Using CasEMBLR | 185 |
| | <i>Tadas Jakóciūnas, Emil D. Jensen, Michael K. Jensen, and Jay D. Keasling</i> | |
| 13 | A Modified Gibson Assembly Method for Cloning Large DNA Fragments with High GC Contents | 203 |
| | <i>Lei Li, Weibong Jiang, and Yinhua Lu</i> | |
| 14 | Coupling Yeast Golden Gate and VEGAS for Efficient Assembly of the Violacein Pathway in <i>Saccharomyces cerevisiae</i> | 211 |
| | <i>James Chuang, Jef D. Boeke, and Leslie A. Mitchell</i> | |
| PART III SCREENING AND SELECTION OF SYNTHETIC METABOLIC PATHWAYS | | |
| 15 | Multi-capillary Column Ion Mobility Spectrometry of Volatile Metabolites for Phenotyping of Microorganisms | 229 |
| | <i>Christoph Halbfeld, Jörg Ingo Baumbach, Lars M. Blank, and Birgitta E. Ebert</i> | |
| 16 | Selection of Highly Expressed Gene Variants in <i>Escherichia coli</i> Using Translationally Coupled Antibiotic Selection Markers | 259 |
| | <i>Maja Rennig, Daniel O. Daley, and Morten H.H. Nørholm</i> | |
| 17 | Design, Engineering, and Characterization of Prokaryotic Ligand-Binding Transcriptional Activators as Biosensors in Yeast | 269 |
| | <i>Francesca Ambri, Tim Snoek, Mette L. Skjoedt, Michael K. Jensen, and Jay D. Keasling</i> | |
| 18 | A Capture-SELEX Strategy for Multiplexed Selection of RNA Aptamers Against Small Molecules | 291 |
| | <i>Lasse H. Lauridsen, Holger B. Doessing, Katherine S. Long, and Alex T. Nielsen</i> | |
| 19 | High-Throughput Microfluidics for the Screening of Yeast Libraries | 307 |
| | <i>Mingtao Huang, Haakan N. Joensson, and Jens Nielsen</i> | |
| 20 | Growth-Coupled Carotenoids Production Using Adaptive Laboratory Evolution | 319 |
| | <i>Luis H. Reyes and Katy C. Kao</i> | |
| PART IV LEARNINGS FROM DESIGN-BUILD-TEST-BASED PROJECTS | | |
| 21 | Two-Scale ¹³ C Metabolic Flux Analysis for Metabolic Engineering | 333 |
| | <i>David Ando and Hector Garcia Martin</i> | |
| | <i>Index</i> | 353 |

Contributors

- FRANCESCA AMBRI • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- DAVID ANDO • *Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; Joint BioEnergy Institute, Emeryville, CA, USA*
- FRANCES H. ARNOLD • *Department of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, USA*
- JÖRG INGO BAUMBACH • *Faculty of Applied Chemistry, Reutlingen University, Reutlingen, Germany*
- LARS M. BLANK • *iAMB—Institute of Applied Microbiology, ABBt—Aachen Biology and Biotechnology, RWTH Aachen University, Aachen, Germany*
- KAI BLIN • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- JEFF D. BOEKE • *School of Medicine, Johns Hopkins University, Baltimore, MD, USA; Institute for Systems Genetics, New York University Langone Medical Center, New York, NY, USA*
- SABINE BRINKMANN-CHEN • *Department of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, USA*
- JACKSON K.B. CAHN • *Department of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, USA*
- MIGUEL A. CAMPODONICO • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- PABLO CARBONELL • *Manchester Centre for Fine and Speciality Chemicals (SYNBIOCHEM), Manchester Institute of Biotechnology, University of Manchester, Manchester, UK*
- DANIEL P. CETNAR • *Department of Chemical Engineering, Pennsylvania State University, University Park, PA, USA*
- JAMES CHUANG • *Department of Biomedical, High Throughput Biology Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA*
- DANIEL O. DALEY • *Department of Biochemistry and Biophysics, Center for Biomembrane Research, Stockholm University, Stockholm, Sweden*
- BAUDOIN DELÉPINE • *iSSB, Genopole, CNRS, UEVE, Université Paris-Saclay, Évry, France; MICALIS Institute, INRA, Jouy-en-Josas Cedex, France*
- HOLGER B. DOESSING • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- BIRGITTA E. EBERT • *iAMB—Institute of Applied Microbiology, ABBt—Aachen Biology and Biotechnology, RWTH Aachen University, Aachen, Germany*
- JEAN-LOUP FAULON • *Manchester Centre for Fine and Speciality Chemicals (SYNBIOCHEM), Manchester Institute of Biotechnology, University of Manchester, Manchester, UK; iSSB, Genopole, CNRS, UEVE, Université Paris-Saclay, Évry, France; MICALIS Institute, INRA, Jouy-en-Josas Cedex, France*
- ADAM M. FEIST • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark; Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA*

- HECTOR GARCIA MARTIN • *Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; Joint BioEnergy Institute, Emeryville, CA, USA*
- CHRISTOPH HALBFELD • *iAMB—Institute of Applied Microbiology, ABBt—Aachen Biology and Biotechnology, RWTH Aachen University, Aachen, Germany*
- SEAN M. HALPER • *Department of Chemical Engineering, Pennsylvania State University, University Park, PA, USA*
- MARKUS J. HERRGÅRD • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- SZU-YI HSU • *Department of Biochemistry, Molecular Biology, and Biophysics and BioTechnology Institute, University of Minnesota, Twin Cities, Saint Paul, MN, USA*
- MINGTAO HUANG • *Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden; Novo Nordisk Foundation Center for Biosustainability, Chalmers University of Technology, Gothenburg, Sweden*
- MOHAMMAD MAZHARUL ISLAM • *Chemical and Biomolecular Engineering, University of Nebraska-Lincoln, Lincoln, NE, USA*
- TADAS JAKOČIUNAS • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- EMIL D. JENSEN • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- MICHAEL K. JENSEN • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- SHEILA INGEMANN JENSEN • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- WEIHONG JIANG • *CAS Ley Laboratory of Synthetic Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, People's Republic of China*
- HAAKAN N. JOENSSON • *Division of Nanobiotechnology and Proteomics, KTH Royal Institute of Technology, Science for Life Laboratory, Solna, Sweden; Novo Nordisk Foundation Center for Biosustainability, KTH Royal Institute of Technology, Stockholm, Sweden*
- KATY C. KAO • *Department of Chemical Engineering, Texas A&M University, College Station, TX, USA*
- JAY D. KEASLING • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark; Joint BioEnergy Institute, Emeryville, CA, USA; Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA, USA; Department of Bioengineering, University of California, Berkeley, CA, USA*
- MICHAEL S. KLAUSEN • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- LASSE H. LAURIDSEN • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- IDA LAURITSEN • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- SANG YUP LEE • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark; Metabolic and Biomolecular Engineering National Research Laboratory, Department of Chemical and Biomolecular Engineering (BK21 Plus Program), Center for Systems and Synthetic Biotechnology, Institute for the BioCentury, BioInformatics Research Center, and BioProcess Engineering Research Center, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea*

- LEI LI • *CAS Ley Laboratory of Synthetic Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, People's Republic of China*
- KATHERINE S. LONG • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- YINHUA LU • *CAS Ley Laboratory of Synthetic Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, People's Republic of China*
- VIRGINIA MARTÍNEZ • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- LESLIE A. MITCHELL • *Institute for Systems Genetics, New York University Langone Medical Center, New York, NY, USA*
- MORTEN H.H. NØRHOLM • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- ALEX TOFTGAARD NIELSEN • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- JENS NIELSEN • *Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden; Novo Nordisk Foundation Center for Biosustainability, Chalmers University of Technology, Gothenburg, Sweden*
- MAJA RENNIG • *Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- LUIS H. REYES • *Process and Product Design Group (GDPP), Department of Chemical Engineering, Universidad de los Andes, Bogotá D.C., Colombia*
- HELENE LUNDE ROBERTSEN • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- CARLOTTA RONDA • *Systems and Synthetic Biology, Columbia University Medical Center, New York, NY, USA*
- RAJIB SAHA • *Chemical and Biomolecular Engineering, University of Nebraska-Lincoln, Lincoln, NE, USA*
- HOWARD M. SALIS • *Department of Chemical Engineering, Pennsylvania State University, University Park, PA, USA; Department of Biological Engineering, Pennsylvania State University, University Park, PA, USA*
- METTE L. SKJOEDT • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- MICHAEL J. SMANSKI • *Department of Biochemistry, Molecular Biology, and Biophysics and BioTechnology Institute, University of Minnesota, Twin Cities, Saint Paul, MN, USA*
- TIM SNOEK • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- MORTEN O.A. SOMMER • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- SUMESH SUKUMARA • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- YAOJUN TONG • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*
- PIER-LUC TREMBLAY • *School of Chemistry, Chemical Engineering and Life Science, Wuhan University of Technology, Wuhan, People's Republic of China; The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark*

TILMANN WEBER • *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark*

TIAN ZHANG • *School of Chemistry, Chemical Engineering and Life Science, Wuhan University of Technology, Wuhan, People's Republic of China; The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark*

Part I

Parts and Models for Designing Synthetic Metabolic Pathways

Chapter 1

Parts Characterization for Tunable Protein Expression

Michael S. Klausen and Morten O.A. Sommer

Abstract

Flow-seq combines flexible genome engineering methods with flow cytometry-based cell sorting and deep DNA sequencing to enable comprehensive interrogation of genotype to phenotype relationships. One application is to study the effect of specific regulatory elements on protein expression. Constructing targeted genomic variation around genomically integrated fluorescent marker genes enables rapid elucidation of the contribution of specific sequence variants to protein expression. Such an approach can be used to characterize the impact of modifications to the Shine-Dalgarno sequence in *Escherichia coli*.

Key words Flow-seq, Flow cytometry cell sorting, Shine-Dalgarno sequence, Next-generation sequencing

1 Introduction

The ability to reliably predict phenotypic changes in response to modified genomic sequences is critical in order to advance the fields of bioengineering and synthetic biology, which increasingly depend on models to predict system behavior. Even small changes in the DNA sequence of regulatory elements, such as the Shine-Dalgarno sequence can lead to substantial phenotypic variation presenting a challenge for predictable engineering of biological systems. Several complementary approaches can address this general issue: Bottom up mechanistic models [1, 2], generation of synthetic regulatory elements [3] and comprehensive empirical characterization [4, 5]. The sensitivity to minor sequence changes in regulatory elements such as the Shine-Dalgarno sequence represents both a challenge and an opportunity. It is a challenge since small changes in sequence composition can lead to unpredicted changes in expression level. However, this is also an opportunity for efficient and multiplexed engineering, if our models can adequately predict these changes.

To tackle this problem and add further to the bioparts engineering toolbox, EMOPEC was developed as a

near-comprehensive database of measured expression levels of the Shine-Dalgarno sequence [5]. The EMOPEC database is based on Flow-seq experiments in which the Shine-Dalgarno sequence upstream of a chromosomally encoded Green Fluorescent Protein (GFP) was randomized and then subjected to sorting based on the resulting GFP expression with a flow cytometer. Each sorted sub-population with a defined GFP expression level range was sequenced to determine which Shine-Dalgarno sequences led to specific expression levels. Based on such approach the GFP expression distribution for individual Shine-Dalgarno sequences can be determined providing reliable estimates for their individual resulting expression strength. The process is shown in Fig. 1.

It is well characterized that the impact of a given Shine-Dalgarno sequence is greatly dependent on its genetic context [6]. Accordingly, great care should be taken when reapplying the measured expression levels into engineering metabolic pathways or synthetic biology circuits, since the ribosome binding site is highly dependent on the local secondary structure of the mRNA. However, if modifications to the Shine-Dalgarno sequence can be done with minimal impact to local secondary structure of the messenger RNA, the relative rank order of the expression strength of specific Shine-Dalgarno sequences is likely maintained. To most effectively ensure this, EMOPEC deploys an algorithm to select Shine-Dalgarno sequences that most effectively explore different expression strengths while minimizing secondary structure changes. Accordingly, it is possible to test a wide range of protein expression strengths while only making minimal changes to the respective Shine-Dalgarno sequence. This allows for parallel and efficient genome editing tuning the expression level of native genes.

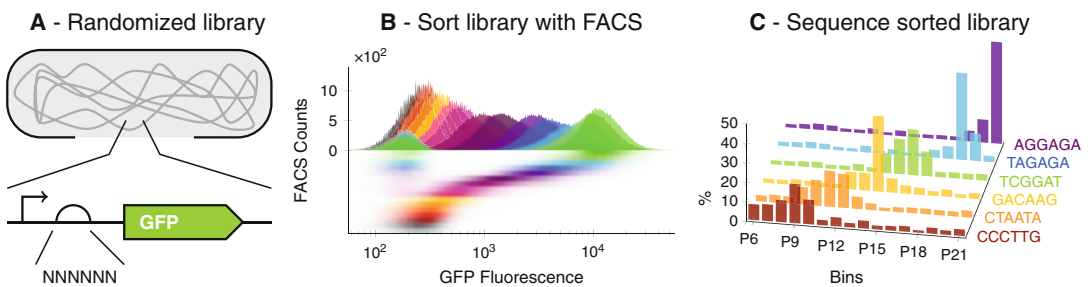


Fig. 1 An overview of the Flow-seq process. (a) A fluorescent reporter gene is integrated downstream of a constitutive promoter and a randomized Shine-Dalgarno sequence. (b) The library is sorted with flow cytometry into several bins of different fluorescent intensity; corresponding to different protein expression levels. (c) The randomized region of each clone in the sorted bins is sequenced and the number of occurrences of each sequence is counted. From the distribution over bins, a weighted expression level can be calculated for tens of thousands of distinct sequences

The Flow-seq approach outlined in this chapter describes how to make and characterize a library of Shine-Dalgarno sequences. However, the method can be deployed for a wide array of parts characterization problems enabling a larger empirical data set that can improve our models of cellular regulation.

2 Materials

2.1 Host Strain Construction

1. Ultrapure H₂O.
2. Q5[®] Hot Start High-Fidelity 2× Master Mix (New England Biolabs).
3. Genomic or plasmid DNA with GFP and a kanamycin resistance gene.
4. Thermocycler for PCR.
5. NucleoSpin[®] Gel and PCR Clean-up kit (Macherey-Nagel).
6. NanoDrop 2000 (Thermo-Fischer).
7. An *E. coli* K12 MG1655 culture.
8. pKD46 plasmid [7].
9. pCP20 plasmid [8].
10. LB-Lennox: 10 g/L tryptone, 5 g/L yeast extract, 5 g/L NaCl.
11. LB-amp: LB-Lennox + 100 µg/mL ampicillin.
12. LB-kan: LB-Lennox + 50 µg/mL kanamycin.
13. Absorption spectrophotometer.
14. LB agar plates.
15. LB-amp agar plates.
16. LB-kan agar plates.
17. 10% stock L-arabinose solution.
18. 30 °C shaking incubator.
19. 37 °C shaking incubator.
20. 42 °C shaking incubator.
21. Centrifuge with cooling.
22. Cuvette for electroporator with 1 mm gap.
23. Electroporator.

2.2 Library and MAGE Oligo Design

1. A computer with a modern browser and a genome editing program.

2.3 Library Construction

1. pMA7 plasmid [9].
2. Ultrapure H₂O.
3. LB-Lennox (*see Note 1*).

4. LB-amp: LB-Lennox + 100 µg/mL ampicillin.
5. LB agar plates.
6. LB-amp agar plates.
7. Absorption spectrophotometer.
8. 10% stock L-arabinose solution.
9. 37 °C shaking incubator.
10. Centrifuge with cooling.
11. Cuvette for electroporator with 1 mm gap.
12. Electroporator.

2.4 Fluorescence-Activated Cell Sorting

1. LB-amp: LB-Lennox + 100 µg/mL ampicillin.
2. 250 mL Shaking flask with baffles.
3. Absorption spectrophotometer.
4. Flow cytometer with sorting capabilities.
5. 5 mL 24-well deep-well multiplate.

2.5 PCR Amplification

1. Genomic DNA Purification Kit (Thermo Fischer Scientific).
2. Q5[®] Hot Start High-Fidelity 2× Master Mix (New England Biolabs).
3. Ultrapure H₂O.
4. Thermocycler for PCR.
5. NucleoSpin[®] Gel and PCR Clean-up (Macherey-Nagel).
6. Bioanalyzer 2100 (Agilent).
7. Qubit Fluorometer (Thermo Fischer Scientific).
8. Qubit dsDNA HS assay kit (Thermo Fischer Scientific).

2.6 Data Analysis

1. A computer with Python installed.

3 Methods

3.1 Host Strain Construction

FLP-FRT recombination is used here to make *E. coli* clones with a chromosomally encoded Green Fluorescent Protein (GFP) reporter. The final construct is shown in Fig. 2a. The accuracy of the Flow-seq method can be improved using a construct with an additional, orthogonal, reporter held at constant expression levels as a normalization of the main reporter levels to account for cell cycle and other factors affecting global expression levels. Such a construct is shown in Fig. 2b. Plasmid pKD46 [7] will be used to recombine the construct into the genome and pCP20 will be used to express the FLP recombinase [8]. Both plasmids are heat sensitive and should be propagated at 30 °C degrees.



Fig. 2 Constructs to make before randomizing the target sequence. (a) The simplest form, where a GFP reporter gene is expressed individually. (b) A dual reporter system, where an additional, orthogonal, and constantly expressed reporter (RFP) is added to normalize the fluorescence from the reporter (GFP) with varied expression levels. Note that the two reporters are on opposite strands to prevent the expression of the upstream gene to affect the expression of the downstream gene

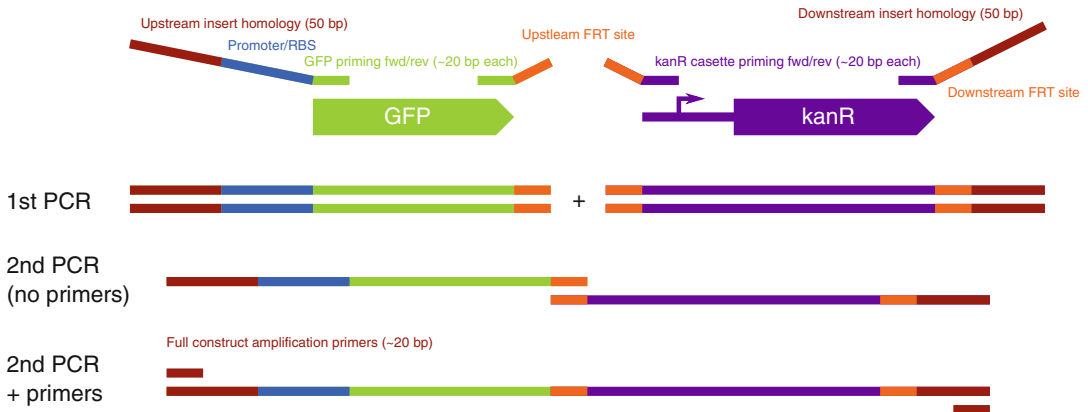


Fig. 3 Primers used to amplify and combine the different genes into the final construct

1. Design the primers needed for amplification of GFP with a strong SD sequence and a constant promoter. BB_J23100 from the BioBricks collection is used here. Similarly, primers for the amplification of kanR with FRT sites. Add an overlap for overlap extension PCR as well. How the primers should be designed can be seen in Fig. 3, as well as a diagram of the different PCR reactions. Good choices for integration site when designing homology arms are either downstream of *glnS*, a highly expressed region, or inside an endogenous reporter (see Note 2).
2. Mix PCR reactions for each fragment:
 - (a) 25 μL Q5[®] Hot Start High-Fidelity 2 \times Master Mix.
 - (b) 2.5 μL 10 μM forward Primer.
 - (c) 2.5 μL 10 μM reverse Primer.
 - (d) 25 ng extracted genomic DNA.
 - (e) H₂O to 50 μL .
3. Put the PCR reaction into the thermocycler and run the following program: 98 $^{\circ}\text{C}$ for 30 s, 30 cycles of: 98 $^{\circ}\text{C}$ for 10 s, 60 $^{\circ}\text{C}$ for 20 s, 72 $^{\circ}\text{C}$ for 10 s. Finally, run for 30 s at 72 $^{\circ}\text{C}$ and hold at 4 $^{\circ}\text{C}$ (see Note 3).

4. Clean up the PCR products using NucleoSpin[®] Gel and PCR Clean-up kit.
5. Measure PCR product concentration using NanoDrop.
6. In one reaction, combine:
 - (a) 25 μ L equimolar amounts of each template, 50–125 ng DNA per template diluted in H₂O.
 - (b) 25 μ L Q5[®] Hot Start High-Fidelity 2 \times Master Mix.
7. Put the PCR reaction into the thermocycler and run the following program: 98 °C for 30 s, 15 cycles of: 98 °C for 10 s, 60 °C for 20 s, 72 °C for 10 s. Finally, run for 30 s at 72 °C and hold at 4 °C.
8. Add 2.5 μ L of each of the final primers (10 μ M), which anneal at either end of the full construct.
9. Run an additional 20 cycles in the thermocycler using identical settings as previous step.
10. Clean up the PCR product with the final construct by running the PCR product on an agarose gel, excising the right band size and purify it with NucleoSpin[®] Gel and PCR Clean-up kit.
11. Grow an *E. coli* K12 MG1655 strain transformed with pKD46 as an overnight culture in LB-amp at 30 °C.
12. Inoculate 100 μ L of the overnight culture in 5 mL LB-amp + 0.2% arabinose and incubate at 30 °C with shaking until OD600 = 0.4.
13. Place the culture on ice for 10 min. Keep the culture cold until electroporation (**step 20**).
14. Centrifuge the culture for 7 min at 7000 $\times g$ in a cooled centrifuge.
15. Discard the supernatant and gently resuspend the cells in 5 mL cold H₂O.
16. Centrifuge the culture again for 7 min at 7000 $\times g$ in a cooled centrifuge.
17. Discard the supernatant and gently resuspend in 1 mL cold H₂O. Transfer cell suspension to a prechilled 1.5 mL tube.
18. Centrifuge the cell suspension for 30 s at 10,000 $\times g$.
19. Discard the supernatant and gently resuspend cells in 100 μ L cold H₂O.
20. Transfer 0.5–1 μ L (maximum 0.5 μ g) of PCR product with the final construct to a 1.5 mL tube.
21. Add 50 μ L washed cells to the same tube. Mix slowly with the tip, do not pipet up and down to mix.
22. Add 50 μ L of cells + PCR product to a prechilled electroporation cuvette with 1 mm gap.

23. Electroporate at 1.8 kV, 200 Ω , 25 μ F, or equivalent *E. coli* electroporator setting.
24. Add 1 mL LB-amp to the cuvette and transfer cells in LB to a new culture tube with 4 mL LB-amp.
25. Incubate the culture at 42 °C with shaking for 1–2 h.
26. Plate 100 μ L 1:10 dilution, 100 μ L and 300 μ L on LB-kan agar plates and incubate overnight at 37 °C.
27. Re-streak colonies and verify loss of pKD46 with PCR and gel electrophoresis (*see steps 2–3* for PCR).
28. Transform the strain with pCP20 (pCP20 confers ampicillin resistance).
29. Inoculate strain with pCP20 into LB and grow at 42 °C overnight to induce FLP and select for loss of pCP20.
30. Dilute the culture 10^{-5} , 10^{-6} , 10^{-7} fold and plate 100 μ L of each dilution on an LB agar plate. Incubate overnight at 30 °C.
31. Re-streak individual colonies on LB-kan, LB-amp, and LB. Be sure to streak LB last, to avoid false positives. Incubate overnight at 37 °C for LB-kan and LB, 30 °C for LB-amp.
32. LB-kan and LB-amp should be clear to indicate successful recombination and loss of pCP20, respectively.

3.2 Library and MAGE Oligo Design

Depending upon goals of the experiment, decide upon either partial or full randomization of the regulatory sequence to be interrogated. Six or seven nucleotides can be completely randomized for a diversity of $4^6 = 4096$ or $4^7 = 16,384$, respectively, but for larger sequences a subsampling of the sequence space is needed. For subsampling of sequence space where there is existing data, Red-Libs [10] can be utilized.

Since MAGE is used to generate the diversity, there will be a bias of strains with the initial sequence. This sequence was constructed to be highly fluorescent to aid in host strain construction; however, an abundance of highly fluorescent cells can bias the sorting. To get a better distribution of data, the fluorescence is first limited by introducing an anti-Shine-Dalgarno sequence. To extend the method to other biological parts, the complement of the consensus sequence or a random nucleotide sequence can be used.

1. Download the *E. coli* K12 MG1655 reference genome (http://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3). Click the “send” button on the top right of the sequence view, check “Complete record,” “File,” and choose “Genbank (full)” as a format.
2. Use a genomic sequence editor, e.g., CLC Main Workbench, to make changes to the genome sequence file reflecting the changes made in the real organism in Subheading 3.1, i.e.,

insert the GFP construct into the genome. Save the changes to a new genbank file.

3. Go to the MODEST [11] webserver (<http://modest.biosustain.dtu.dk>). In “Step 1: Choose an organism,” select “other,,” and upload the newly created genbank file. Leave the other options as default and press “Create genome cfg.”
4. In “Step 2,,” select “genome” under “Gene,,” and select “find_mutation” under “Operation.” Paste in “TAGAGATTAAAG [AGGAGA->NNNNNN]AATACTAGATGA” into the mutation field. The part before the square bracket is the sequence upstream of the target SD, “AGGAGA” is the current SD, and “->NNNNNN” means to mutate the SD into a fully randomized sequence. The remaining part is the downstream sequence, after the SD. Press “Add line.”
5. Repeat **step 4** using the input “TAGAGATTAAAG[AGGAGA->TCCTCC]AATACTAGATGA” to make the anti-SD mutation oligo.
6. Create the oligos by clicking “Submit” and wait for the job to finish.

3.3 Library Construction

Nucleotide oligos for MAGE are obtained from Integrated DNA Technologies as the library is completely randomized and can thus be encoded on a single degenerate oligo. As an alternative, microarray-based nucleotide synthesis can be used to create libraries of arbitrary composition [12].

1. Grow the *E. coli* strain from Subheading 3.1 transformed with the pMA7 plasmid [9] to OD₆₀₀ = 0.4 in 5 mL LB-amp at 37 °C with shaking (*see Note 4*).
2. To induce *beta* and *dam* expression, add L-arabinose to a final concentration of 0.2% by adding 100 μL 10% L-arabinose stock to 5 mL LB-amp medium.
3. Incubate an additional 15 min at 37 °C with shaking.
4. Place cell culture on ice and cool for at 15 min. Keep cold (<4 °C) from this point until electroporation in **step 14**.
5. Centrifuge the culture for 7 min at 7000 × *g* in a cooled centrifuge.
6. Discard the supernatant and gently resuspend the cells in 5 mL cold H₂O.
7. Centrifuge the culture again for 7 min at 7000 × *g* in a cooled centrifuge.
8. Discard the supernatant and gently resuspend in 1 mL cold H₂O. Transfer cell suspension to a prechilled 1.5 mL tube.
9. Centrifuge the cell suspension for 30 s at 10,000 × *g*.

10. Discard the supernatant and gently resuspend cells in 100 μL cold H_2O .
11. Transfer 1 μL of 10 μM anti-SD oligo to a prechilled 1.5 mL tube.
12. Add 50 μL washed cells to the same tube. Mix slowly with the tip, do not pipet up and down to mix.
13. Add the 51 μL of cells + anti-SD oligo to a prechilled electroporation cuvette with 1 mm gap.
14. Electroporate at 1.8 kV, 200 Ω , 25 μF , or equivalent *E. coli* electroporator setting.
15. Add 1 mL LB-amp to the cuvette and transfer cells in LB to a new culture tube with 4 mL LB-amp.
16. Incubate the culture overnight at 37 $^\circ\text{C}$ with shaking (*see Note 5*).
17. Dilute the culture 10^{-5} , 10^{-6} , 10^{-7} fold and plate 100 μL of each dilution on an LB-amp agar plate.
18. Incubate overnight at 37 $^\circ\text{C}$.
19. Look at the agar plates under blue light and re-streak single colonies that are not fluorescing to new LB-amp agar plates (*see Note 6*).
20. Repeat **steps 1–15** with the randomized oligo for 6 cycles. When repeating the MAGE cycles, electroporated cells can either be grown directly to log-phase (restart from **step 2**) or grown overnight (restart from **step 1**) depending on convenience.

3.4 Fluorescence-Activated Cell Sorting

1. Inoculate 100 μL from the final library into 5 mL LB-amp and incubate overnight at 37 $^\circ\text{C}$.
2. From the overnight culture, inoculate 100 μL into 50 mL LB-amp in a 250 mL shaking flask with baffles.
3. Incubate at 37 $^\circ\text{C}$ at 250 rpm or higher shaking to make sure there is aeration in the media.
4. When the culture is at OD600 between 0.4 and 0.6, place the culture on ice.
5. Set up the flow cytometer with the culture.
6. Adjust gate settings so cells are sorted into 16 evenly spaced bins.
7. Sort cells into 100 μL LB-amp media.
8. Regrow each sorted bin overnight in 1.5 mL LB-amp at 37 $^\circ\text{C}$ in a deep-well 24-well multiplate.
9. Take out 500 μL from each culture and store as a glycerol stock.

3.5 PCR Amplification

1. Extract genomic DNA from each overnight culture (from Sub-heading 3.4, **step 8**) using Genomic DNA Purification Kit (Thermo Fischer).
2. Design primer to amplify an area around the engineered diversity (*see* **Note 7**).
3. Set up the following PCR reaction, one for each bin:
 - (a) 25 μL Q5[®] Hot Start High-Fidelity 2 \times Master Mix.
 - (b) 2.5 μL 10 μM forward Primer.
 - (c) 2.5 μL 10 μM reverse Primer.
 - (d) 1 μg extracted genomic DNA.
 - (e) H₂O to 50 μL .
4. Put the PCR reaction into the thermocycler and run under the following conditions: 98 °C for 30 s, 30 cycles of: 98 °C for 10 s, 60 °C for 20 s, 72 °C for 10 s. Finally, run for 30 s at 72 °C and hold at 4 °C.
5. Clean up the PCR products using NucleoSpin[®] Gel and PCR Clean-up.
6. Verify the size of the PCR product using on a Bioanalyzer 2100 or using gel electrophoresis. Determine the concentration of PCR product with Qubit.

3.6 Data Analysis

PCR products should be sequenced with a short read high-throughput sequencing method, e.g., Illumina MiSeq yielding 10–20 million paired-end reads. Scripts implemented in Python as well as a demonstration of the data are available in <https://github.com/micked/flowseq-scripts>.

1. Count occurrences of each SD sequence in each bin. This can be accomplished using Python and regular expressions.
2. Filter away sequences with less than a total of 50 reads.

Obtain the mean fluorescence of each flow cytometry bin by fitting a Gaussian curve over the histogram of fluorescence activations for each bin. An example bin with a fitted curve is shown in Fig. 4.
3. Use the equation from [13] to calculate the weighted mean expression level: $f_s = (\sum_b n_{b,s} / n_b \times e_b) / (\sum_b n_{b,s} / n_b)$ where e_b is the mean value of the Gaussian fit mentioned above, n_b is the total count of all SD sequences in bin b , and $n_{b,s}$ is the count of SD sequence s in bin b .
4. Perform **step 4** on all the sequences that passed the filter in **step 2**. This will result in the final table of expression values.

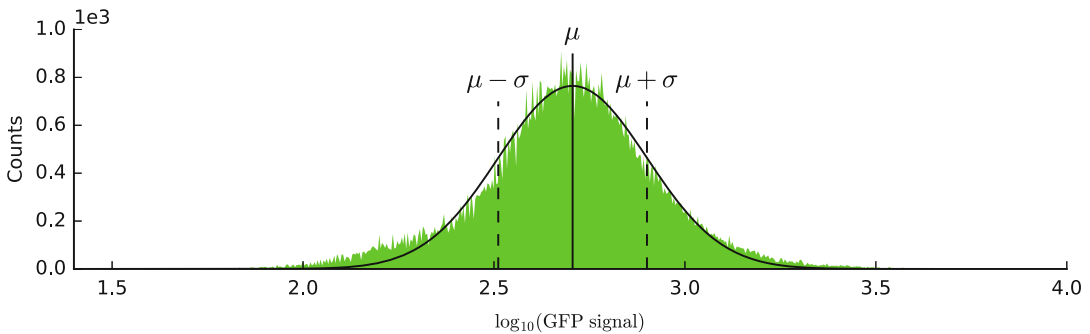


Fig. 4 Histogram of fluorescence activations of a single bin, with the mean (μ) and standard variation (σ)

4 Notes

1. The salt concentration of the LB media is very important for MAGE and recombineering. High salt concentration can lead to extremely low allelic replacement frequencies.
2. By integrating the construct in the middle of an endogenous reporter in a way that knocks out the reporter, the antibiotic selection and pCP20 recombination steps can be skipped. Mutants are found by screening on agar plates corresponding to the chosen reporter. For example, the *lacZ* gene can be used for blue-white screening or *galK* or *malK* for red-white screening (MacConkey agar with galactose or maltose, respectively, as carbon source). Be aware that integration efficiencies vary, to a point where screening on agar plates is not feasible. Antibiotic selection is thus the safer choice.
3. PCR conditions, especially the 60 °C annealing temperature should be altered to suit the primers and the polymerase.
4. We find that ODs between 0.2 and 0.4 works best. We usually inoculate 100 μL overnight *E. coli* culture in transparent culture tubes and use a spectrophotometer that directly measures culture tubes.
5. It can be tempting to skip this step and grow the cells to near-saturation and plate directly. This is not recommended, since the cells are still recombining with the oligos 3+ h after electroporation.
6. Verify the mutation is the right one either by allele-specific PCR or, preferably, by sequencing.
7. This is usually a simple primer design job. We prefer using Primer3 [14]. Design your primers according to the selected sequencing method, when designing your primers for Illumina 150 bp paired end sequencing, design the primers so the target region is in the middle of a PCR product that is as close to 150 bp as possible.

References

1. Espah Borujeni A, Channarasappa AS, Salis HM (2014) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res* 42:2646–2659. doi:[10.1093/nar/gkt1139](https://doi.org/10.1093/nar/gkt1139)
2. Seo SW, Yang J-S, Kim I, Yang J, Min BE, Kim S et al (2013) Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency. *Metab Eng Elsevier* 15:67–74. doi:[10.1016/j.ymben.2012.10.006](https://doi.org/10.1016/j.ymben.2012.10.006)
3. Mutalik VK, Guimaraes JC, Cambray G, Lam C, Christoffersen MJ, Mai Q-A et al (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat Methods* 10:354–360. doi:[10.1038/nmeth.2404](https://doi.org/10.1038/nmeth.2404)
4. Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y, Arkin AP et al (2013) Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proc Natl Acad Sci U S A* 110:14024–14029. doi:[10.1073/pnas.1301301110](https://doi.org/10.1073/pnas.1301301110)
5. Bonde MT, Pedersen M, Klausen MS, Jensen SI, Wulff T, Harrison S et al (2016) Predictable tuning of protein expression in bacteria. *Nat Methods* 13:233–236. doi:[10.1038/nmeth.3727](https://doi.org/10.1038/nmeth.3727)
6. Salis HM, Mirsky EA, Voigt CA (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* 27:946–950. doi:[10.1038/nbt.1568](https://doi.org/10.1038/nbt.1568)
7. Datsenko KA, Wanner BL (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. doi:[10.1073/pnas.120163297](https://doi.org/10.1073/pnas.120163297)
8. Cherepanov PP, Wackernagel W (1995) Gene disruption in *Escherichia coli*: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. *Gene* 158:9–14. doi:[10.1016/0378-1119\(95\)00193-A](https://doi.org/10.1016/0378-1119(95)00193-A)
9. Lennen RM, Nilsson Wallin AI, Pedersen M, Bonde M, Luo H, Herrgård MJ et al (2016) Transient overexpression of DNA adenine methylase enables efficient and mobile genome engineering with reduced off-target effects. *Nucleic Acids Res* 44:e36. doi:[10.1093/nar/gkv1090](https://doi.org/10.1093/nar/gkv1090)
10. Jeschek M, Gerngross D, Panke S (2016) Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort. *Nat Commun* 7:11163. doi:[10.1038/ncomms11163](https://doi.org/10.1038/ncomms11163)
11. Bonde MT, Klausen MS, Anderson MV, Wallin AIN, Wang HH, Sommer MOA (2014) MODEST: a web-based design tool for oligonucleotide-mediated genome engineering and recombineering. *Nucleic Acids Res* 42:W408–W415. doi:[10.1093/nar/gku428](https://doi.org/10.1093/nar/gku428)
12. Bonde MT, Kosuri S, Gence HJ, Sarup-Lytzen K, Church GM, Sommer MOA et al (2015) Direct mutagenesis of thousands of genomic targets using microarray-derived oligonucleotides. *ACS Synth Biol* 4:17–22. doi:[10.1021/sb5001565](https://doi.org/10.1021/sb5001565)
13. Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D et al (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* 30:521–530. doi:[10.1038/nbt.2205](https://doi.org/10.1038/nbt.2205)
14. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M et al (2012) Primer3—new capabilities and interfaces. *Nucleic Acids Res* 40:e115–e115. doi:[10.1093/nar/gks596](https://doi.org/10.1093/nar/gks596)

Chapter 2

Enzyme Nicotinamide Cofactor Specificity Reversal Guided by Automated Structural Analysis and Library Design

Jackson K.B. Cahn, Sabine Brinkmann-Chen, and Frances H. Arnold

Abstract

The specificity of enzymes for nicotinamide adenine dinucleotide (NAD) or nicotinamide adenine dinucleotide phosphate (NADP) as redox carriers can pose a significant hurdle for metabolic engineering and synthetic biology applications, where switching the specificity might be beneficial. We have developed an easy-to-use computational tool (CSR-SALAD) for the design of mutant libraries to simplify the process of reversing the cofactor specificity of an enzyme. Here, we describe the optimal use of this tool and present methods for its application in a laboratory setting.

Key words Cofactor specificity, Cofactor switch, Protein engineering, Library design, Oxidoreductases

1 Introduction

The need to balance the production and consumption of the hydride transport cofactors nicotinamide adenine dinucleotide hydride (NADH) and nicotinamide adenine dinucleotide phosphate hydride (NADPH) is a common hurdle in the design of efficient synthetic metabolic pathways [1]. Although metabolic engineering approaches have been used to increase the availability of nicotinamide redox equivalents, the best results have come from the targeted engineering of enzymes to alter their cofactor specificities and create a balance between the cofactors produced and consumed in an engineered pathway [2, 3].

Since the first reported engineered specificity switch in 1990 by Scrutton and coworkers [4], altering nicotinamide cofactor specificity has frequently been a target of protein engineering efforts. However, the engineering of these proteins has occurred on an individual, piecemeal basis with no single approach having proven consistently effective. Furthermore, such engineering has remained the province of experts rather than a tool for end-users such as metabolic engineers, due to a requirement for specialized

knowledge or software or the equipment for high-throughput screening of enzyme variants.

To address this, we recently developed a method for automated design of mutant libraries for NAD(P)-dependent oxidoreductases: Cofactor Specificity Reversal–Structural Analysis and Library Design (CSR-SALAD) [5]. This method was validated experimentally by reversing the specificity of four NADP-dependent enzymes and further validated in silico for an additional seven NADP-dependent and 11 NAD-dependent enzymes for which the kinetics of engineered mutant enzymes had been reported in the literature. This approach was designed to be accessible for users outside the field of protein engineering, requiring very little theoretical knowledge or advanced experimental techniques to implement.

As shown in Fig. 1, the approach consists of three steps, laid out in greater detail in the sections below. Firstly, a crystal structure of the protein, bound to its natively preferred cofactor, is uploaded to the CSR-SALAD webserver (<http://cheme.che.caltech.edu/groups/fha/CSRSALAD/index.html>) and analyzed. The webserver then suggests a single specificity-reversal library with a size

CSR-SALAD cofactor specificity reversal structural analysis & library design (a)

Input **About**

Welcome to CSR-SALAD, the Arnold Lab tool for nicotinamide cofactor specificity reversal. You can either upload your own structure (as a .pdb file) OR analyze a published structure by entering an RCSB PDB ID. For more information, please consult the [documentation](#) or [contact us](#) with further questions.

Upload a Protein File Select a .pdb file from your computer: No file selected. **OR** **Enter a PDB ID** Protein PDB ID (e.g., 4TSK):

Maximum library size (plan to screen 2-3x library size clones)

Advanced Options ...

Email Address:

© 2015 California Institute of Technology Citation · Contact · Funding · Terms · Change Log

Analysis Results for 1PIW (b)

| Residue | Type | Codon | AAs |
|---------|-----------|-------|----------|
| SER 210 | Simple | RFC | DSNS |
| ARG 211 | Face | YNC | CRHLPRSY |
| LYS 215 | Didentate | RFK | DECKNRIS |

Suggested library size: 256

The following residues should be targeted first for activity-recovery by site-saturation mutagenesis (grouped by priority):

High priority:
LEU 188

Medium priority:
ALA 251
SER 252
SER 253
THR 255
ASP 256

Low priority:
ARG 214
ARG 216
TYR 227

Analyze (c)

- Detect residues that determine cofactor specificity
- Classify based on position and orientation

Switch

- Design small degenerate codon library for reversal of specificity

Recover

- Identify structural hotspots likely to harbor compensatory mutations

Fig. 1 The web interface of CSR-SALAD (a), and an example of the output (b). CSR-SALAD performs three tasks (c): structure analysis, design of cofactor-switching libraries, and identification of positions for activity recovery. Figure adapted with permission from Cahn et al. [5] Copyright 2016 American Chemical Society

below a threshold specified by the user. After screening of this library, the third step is to search for mutations that recover activity on the new cofactor at predicted “hotspots.” More details of the workings of the CSR-SALAD algorithm are available in the original paper [5].

2 Materials

2.1 Equipment

The equipment listed below assumes that a standard set of instrumentation usually present in a molecular biology/microbiology laboratory, such as pipettes, PCR thermocyclers, DNA gel documentation station, centrifuges, incubators, and autoclave, is accessible.

1. Multichannel pipettes, preferably electronic models.
2. Electroporator (e.g., GenePulser Xcell from Bio-Rad).
3. Shaking incubators for 96-well plates with humidity control.
4. Plate reader(s) with optional fluorescence function.
5. Rotors for centrifuging 96-well plates.
6. 96-well deep-well plates.
7. 96-well assay plates.
8. (Recommended) Liquid handling robot (for example Hamilton Nimbus 2000 or a Tecan model).
9. (Recommended) Biosafety cabinet for colony picking and any kind of sterile work.

2.2 Library Construction

1. Expression plasmid (such as pET22b(+)) containing the gene encoding the protein of interest.
2. Luria Bertani (LB) agar plates supplemented with the antibiotic appropriate for the expression vector. We use ampicillin and refer to these plates as LB_{amp} agar plates.
3. LB liquid medium: 1% Bacto tryptone, 0.5% yeast extract, and 1% NaCl in deionized water. For plates, 1.5% agar is added.
4. LB_{amp}: LB liquid medium + 100 µg/mL ampicillin.
5. SOC (super optimal broth with catabolite repression) medium: 0.5% Yeast extract, 2% tryptone, 10 mM NaCl, 2.5 mM KCl, 10 mM MgCl₂, 10 mM MgSO₄, 20 mM glucose. Autoclave the solution without glucose. Sterilize the glucose solution by passing it through a 0.2 µm filter. Mix the two solutions together.
6. Phusion™ polymerase or any other high fidelity polymerase.
7. PCR-grade water.

8. dNTPs (dATP, dTTP, dCTP, dGTP, 10 mM each). Store at -20°C .
9. 1% Tris-acetate-EDTA (TAE): 40 mM Tris-acetate, 1 mM EDTA.
10. Electrocompetent *Escherichia coli* BL21 E. cloni cells or any other electrocompetent host cell of choice.
11. Miniprep kits (for example from Qiagen or Zymo Research).
12. Agarose gel extraction kit (for example from Promega or Zymo Research).
13. PCR clean up and concentrator kit (Zymo Research).
14. Gibson ligation master-mix.

2.3 Assay Development, Library Screening and Rescreening

1. Sterile tooth picks.
2. Sterile 96-well deep-well plates.
3. Lids (sprayed with 70% ethanol) or AirPore tape (Qiagen) to seal the deep-well plates.
4. Expression media such as LB_{amp} or Terrific Broth (TB_{amp}).
5. Isopropyl β -D-1-thiogalactopyranoside (IPTG).
6. Lysis buffer (*see* Subheading 3.4).
7. Assay buffers containing NADH and NADPH (*see* Subheading 3.4).
8. LB_{amp} agar plates.

3 Methods

3.1 Structure Preparation

3.1.1 Existing Structures

For many proteins, crystal structures of the enzyme bound to its natively preferred cofactor can be found on the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) [6]. For these structures, no modification is likely required. These structures can be accessed using their accession codes or downloaded from the PDB and uploaded to CSR-SALAD (*see* Subheading 3.2, **step 1**).

3.1.2 Homology Modeling

If a crystal structure of the protein is available, but without cocrystallized cofactor, the cofactor can often be placed by homology.

1. Go to the PDB and search for the sequence of the protein of interest. Locate the protein structure with the desired cofactor bound which has the highest sequence identity to your protein of interest (*see* **Note 1**). Download both this structure and the apo structure of the protein of interest.

2. Using a structure visualization/manipulation tool such as PyMol [7], align the structure of the homologue to the apo structure and save it.
3. Using a text editor, copy the lines of the PDB file representing the cofactor molecule and paste it into the file of the apo structure of your protein.
4. Open this combined file again in the structure visualization tool. If necessary, adjust the rotamers for residues around the cofactor to eliminate obvious steric clashes and provide reasonable hydrogen bonding patterns.

If no structure is available, online tools for homology modeling, such as SWISS-MODEL [8] or Phyre2 [9] can create models based on existing structures. After creation, **steps 1–4** above can be used to place the cofactor.

3.2 Library Design

1. Navigate to the CSR-SALAD website (<http://cheme.che.caltech.edu/groups/fha/CSRSALAD/index.html>). Structures can be accessed by their 4-character PDB accession code or uploaded using the “Browse” button (Fig. 1) (*see Note 2*).
2. Select a maximum library size. This library size should be at most half the possible throughput of your experimental setup as determined by equipment for handling 96-well plates—shaker-incubators, centrifuges, liquid handlers, and automation. Library sizes below 40 will be rejected, and library sizes above 10^5 may be less straightforward to clone.
3. CSR-SALAD allows for the selection of advanced options, but the use of these is not recommended. For details, consult the CSR-SALAD documentation, Section 2.4.
4. Optionally, enter your email address.
5. Click “Analyze.” Wait a few seconds while calculations are completed.
6. Take note of the cofactor-switching library, at the top, and the suggested saturation mutagenesis positions for activity recovery, below (Fig. 1).

3.3 Library Construction

Any number of methods for primer design and mutagenic cloning can be employed for the generation of mutant libraries. Here we present a modified “exponential” QuikChange protocol developed in our group and first published by Engqvist et al. [10] This method is suitable for all PCR steps described in this chapter (*see Notes 3 and 4*).

1. Primer design is shown in Fig. 2. Each primer has a complementary 20-base 5′ end and a 3′ nonoverlapping component with a melting temperature of approximately 55 °C as determined with the Nearest Neighbor method [11]. The

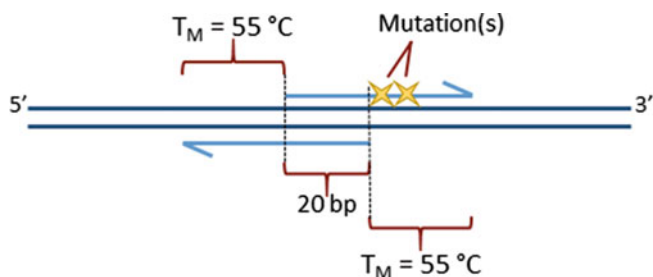


Fig. 2 Primer design approach for library construction. This overlapping design with mutations only on one primer allows for exponential amplification of the full plasmid, eliminating the need for ligation to a separate vector while providing homologous regions for plasmid circularization by Gibson ligation

degenerate mutagenic codon(s) should fall in the first 10–15 bases of only one of these primers, while the other primer should be exactly complementary to the template. Primers can be obtained from any oligonucleotide synthesis service that allows for mixed bases and should be prepared according to manufacturer's instructions to a final concentration of 10 μM .

2. PCR mix: 10 μL NEB HF buffer (5 \times), 1 μL dNTPs (10 μM each), 1 μL of each primer, 0.5 μL plasmid template (~100 ng/ μL), 0–4 μL DMSO (*see Note 5*), 0.5 μL PhusionTM polymerase (*see Note 6*), and PCR-grade water up to 50 μL .
3. PCR protocol: 98 °C for 30 s; 98 °C for 10 s, 55 °C for 10 s, 72 °C for 30 s/kb (25 cycles); 72 °C for 10 min; 10 °C hold step (not required).
4. Add 1 μL *DpnI* and incubate at 37 °C for 3 h.
5. Add 10 μL 6 \times DNA loading dye. Run samples on preparative TAE agarose gel. Check for correctly sized products. Excise on UV or blue light table and extract using gel extraction kits from e.g., Promega or Zymo Research, eluting in 30 μL of PCR-grade water or elution buffer.
6. Add 5 μL eluent to 15 μL 4/3 \times Gibson ligation [12] mix (*see Note 7*): Incubate at 50 °C for 1 h.
7. Purify Gibson ligation product using a Zymo Research Clean & Concentrator-5 kit or other low-volume PCR cleanup kit, eluting in 7–10 μL PCR-grade water or elution buffer (*see Note 8*).
8. Add 1 μL of the ligated plasmid DNA to 50 μL of electrocompetent BL21(DE3) *E. coli* cells. Transfer to ice-cold electroporation cuvette. Electroporate and add 700 μL of SOC media and incubate with shaking at 37 °C for 45 min (*see Note 9*).

9. On the same day, transform 50 μL of electrocompetent cells each with the parent plasmid (serves as positive control) and the vector (serves as negative control).
10. Plate varying volumes of the recovery mixture on LB_{amp} agar plates. Spread evenly, allow to dry in a sterile location, and incubate upside-down at 37 °C overnight (*see Note 10*).
11. Pick 5–10 colonies of the library into 4-mL LB_{amp} cultures and grow overnight at 37 °C with shaking. Extract the plasmid DNA with a miniprep kit of choice and sequence with appropriate primers to verify successful library construction.

3.4 Assay Development

Every protein behaves in a slightly different fashion. Before beginning a screening process, it is crucial to spend time on the assay development since the entire process hinges on having a robust assay. For this, we test the expression and activity of the wild-type enzyme in a 96-well format by growing and “screening” plates (as Subheading 3.5) comprising the parent protein as well as the standard negative controls. This serves two purposes: it ensures that expression and activity are sufficient to be able to detect improvements and also gives a measure of the experimental uncertainty in the assay. Determining the coefficient of variance (CV, the ratio of standard deviation to mean) of the assay is important for setting a screening threshold to minimize false positives, as described in Subheading 3.6.

In the subsequent sections, we present the most common values we use. Parameters marked with an asterisk (*) are those that can be varied to optimize expression or the activity assay. In this section, we discuss some of the common areas for assay improvement.

1. Design an assay buffer. If previous characterization of the protein has been done, information from the literature or from personal experience can be used to determine factors such as optimal pH, ionic strength, and whether there is a need for reducing agents such as DTT. The reaction substrate concentration should be 2–3 times greater than the K_M for the substrate (*see Note 11*). Make two assay buffers, one for each cofactor, using the same concentration of cofactor for both. This concentration should be 2–3 times the wild-type K_M of the naturally preferred cofactor. If no previous information is available, expression of protein at scale can be used for the determination of K_M s by titration.
2. The simplest parameter to vary is the setup of the assay itself, that is, the ratio of the lysate to assay buffer. In general, the accuracy of these high throughput pipetting steps benefits from the usage of volumes larger than 20 μL , and the total volume per well should be 200 μL with no more than half of that being

lysate. Test-screen different lysate volumes and record the change in absorbance with the plate reader. The optimal lysate volume is the one that when doubled will yield a twofold increase of the signal. If doubling the lysate volume results in less than doubling of the signal, it will be very difficult to distinguish any improved variants in the actual screen. If cofactor is depleted too rapidly to observe, dilution with buffer (not assay or lysis buffer) can make enzymatic activity easier to measure.

3. If activity is too low to measure, or if well-to-well variation is too high ($CV > 15\%$), it may be necessary to optimize expression conditions to increase the signal. Common parameters to alter include medium composition (LB, TB, M9, etc.), expression temperature and time, and IPTG induction concentration and induction time. For particularly challenging proteins, construct design and expression strain selection can also be investigated to achieve higher levels of expression. Depending on experimental setup, expression can be tested using the assay developed in the previous steps or by SDS-PAGE and optionally Western blotting.
4. Frequently the apparent expression level of protein can be negatively impacted by inefficient or overly harsh lysis conditions, so it may be necessary to optimize the composition of the lysis buffer. The buffer and its pH are dependent on the enzyme's preferences and should match those of the assay buffer. The buffer should always contain 10 mg/L DNaseI, and at least 750 mg/L of lysozyme. Common optional components include 2 mM $MgCl_2$ (to enhance DNase activity), 1–5 mM dithiothreitol (DTT), 5–10% glycerol, or a protein extraction reagent such as BugBuster.

3.5 Screening

1. In a sterile environment, use an electronic multichannel pipette to fill a sterile set of 96-well deep-well plates with 300 μ L of LB_{amp}^* .
2. Use sterile tooth picks to transfer single colonies of the library and the control plates into the wells. To ensure screening of the library to 95% completion, aim for threefold oversampling when picking colonies. Each plate should always contain at least four parent colonies, three negative control colonies, and one well that remains un-inoculated for sterility control.
3. Cover the plates with either AirPore tape or lids (spray lids with 70% ethanol and let them dry in the biosafety cabinet to ensure sterility).

4. Grow overnight cultures in an incubator at 37 °C, 80% humidity, and shaking at, for example, 250 rpm (in a shaker with a 50 mm radius).
5. On the next morning, fill another set of 96-well deep-well plates with 600 μL^* of expression medium.
6. Transfer 50 μL^* of the overnight cultures into the fresh expression plates. Using a liquid handling robot is very helpful for this step. Make sure to sterilize or replace tips between each plate.
7. Place the overnight culture plates into a refrigerator for later usage (*see* Subheading 3.6).
8. Cover the freshly inoculated expression plates with sterile lids or tape and continue to grow them in the incubator as described in **step 4**.
9. After 3.5 h, remove plates from incubator and chill them on ice for approximately 20 min (*see* **Note 12**).
10. In the meantime, turn off the humidity control and cool down the shaker to expression temperature, i.e., 20 °C*.
11. Induce the expression by addition of IPTG to a final concentration of 500 μM^* .
12. Continue to grow the cultures at the desired expression temperature for another 20 h*.
13. Centrifuge plates (*see* **Note 13**), discard supernatant, and freeze the pellets in the plates overnight at $-20\text{ }^\circ\text{C}$ (*see* **Note 14**). If desired, plates can be frozen for up to 1 month, although overnight cultures (**step 7**) should not be kept more than 1 week and will need to be converted to glycerol stocks if longer storage is required.
14. On screening day, thaw the plates at room temperature* for 20 min*.
15. Add 250 μL^* of lysis buffer, vortex, and incubate at 37 °C for 1 h*. If an extraction reagent is used, alter lysis time and temperature according to manufacturer's instructions.
16. To pellet the cell debris, centrifuge at $5000 \times g$ for a minimum of 10 min at 4 °C.
17. Transfer appropriate volume of lysate, as determined in the assay development phase, into one or more standard clear bottom assay plates (i.e., from Greiner Bio-One).
18. Ensure that the wells with the lysate do not contain any air bubbles. A quick ($\sim 20\text{ s}$ at $2000 \times g$) centrifugation step can help remove them if need be.
19. Use a multichannel pipette to add the assay buffer (*see* **Note 15**) to the lysate in the assay plate wells. Try to fill all wells of a

single plate as rapidly as possible and promptly place the plate into the plate reader.

20. Record the consumption of NADH or NADPH in each well by measuring absorbance at 340 nm in the plate reader over 3 min. Ideally consumption should be linear over this time span.

3.6 Rescreening

After selecting the best variants from the screen, their improvements in activity need to be validated in a rescreening procedure. This minimizes the risk of two common problems encountered during screening: erroneous individual measurements (false positives) and mixed colonies.

1. Returning to the overnight culture plates that were stored in the refrigerator, streak cells of the hits on LB_{amp} agar plates such that single colonies are obtained for each candidate.
2. Incubate LB_{amp} agar plates at 37 °C overnight.
3. On the next day, fill as many 96-well deep-well plates as needed with 300 µL of expression medium as described above (*see* Subheading 3.5, **step 1**).
4. Pick at least four single colonies of each variant.
5. Include parent and negative control on the rescreen plate(s) as well.
6. Repeat **steps 4–20** from Subheading 3.5.
7. Select the best variants (*see* **Note 16**) for DNA sequencing.
8. If desired, express sequence-verified variants on larger scale (*see* **Note 17**) for purification and thorough biochemical and kinetic characterization.

4 Notes

1. In some cases, it may be better to choose a file with higher sequence identity in the cofactor-binding region over a sequence with higher overall identity.
2. For details on the processing and requirements for uploaded PDB files, consult the CSR-SALAD documentation, available from <http://cheme.che.caltech.edu/groups/fha/CSRSALAD/About.html>, Section 2.2, **item 1**.
3. For the majority of first-round libraries generated by CSR-SALAD, all mutated sites are sufficiently close in primary sequence space to be included on a single primer. In cases where one or more sites are separate, the smaller sub-libraries should be cloned first and isolated plasmid DNA from a bulk

culture of that library can be used as template for subsequent PCRs.

4. When second-round sites are close together, a single common reverse primer can often be used.
5. For many reactions, a standard concentration of 4% DMSO (2 μ L) is sufficient. If the long-range PCR is unsuccessful, a set of four additional PCRs can be attempted with 0, 1, 3, and 4 μ L of DMSO to find a preferred concentration.
6. Polymerase should be added last, and thermocycling initiated as quickly thereafter as possible.
7. Gibson mixes prepared by researchers are often $4/3\times$, while commercial mixes are often $2\times$. To use a $2\times$ mix, simply combine 10 μ L of the mix with 5 μ L of PCR-grade water.
8. Products of this ligation will not be at sufficient concentration to visualize on an agarose gel.
9. Different antibiotics require different recovery times. If using antibiotics other than ampicillin, consult established protocols.
10. The volume of cells to be plated to ensure single colonies is dependent on the competence of the cells and the efficiency of the Gibson ligations. Often it is helpful to plate a range of volumes to determine the optimal quantity, then to repeat the transformation the following night plating several plates at this level.
11. We find this concentration ensures high (and therefore easy-to-measure) activity while also ensuring that significant changes in substrate K_M are reflected in changes to the activity.
12. For our purposes, 3.5 h has been an excellent length of pre-induction growth time. Chilling on ice immediately before induction may help increase the yield of the soluble fraction while reducing the amount of inclusion bodies formed.
13. We found the ideal settings for pelleting the cultures to be $5000 \times g$ for 3 min. If the cells are being pelleted too hard or too long, the subsequent resuspension step will be difficult.
14. To achieve even and efficient lysis, it is imperative that the plates be frozen overnight. A shorter -80°C -freezer step does not replace the longer overnight freezing step.
15. Adding all required components of the assay as one reaction master mix to the lysate reduces the amount of pipetting steps to one, thereby reducing the amount of error.
16. A common error in the analysis of the rescreening data is to average the four individual data points per variant. The goals of the rescreen were to isolate four individual colonies to ensure that the screening results are from monoclonal colonies instead of mixtures and to verify the improved variants. Oftentimes,

two completely different activity levels are observed for one variant in the rescreen, reflecting two different geno-/phenotypes present in the original screen. In that case, only the one that exceeds the activity of the parent is the true hit.

17. “Large scale” is highly individual. For some enzymes, a 50-mL culture will yield high enough enzyme concentrations to collect all the biochemical data required; for other enzymes, 1-L or even larger culture volumes are needed.

Acknowledgements

This work was supported by the Gordon and Betty Moore Foundation through grant number GBMF2809 to the Caltech Programmable Molecular Technology Initiative and by American Recovery and Reinvestment Act (ARRA) funds through the National Institutes of Health Shared Instrumentation Grant Program, grant number S10RR027203 to F.H.A. J.K.B.C acknowledges the support of the Resnick Sustainability Institute (Caltech). The cloning method described in Subheading 3.3 was developed by Martin Engqvist.

References

1. Wang Y, San K-Y, Bennett GN (2013) Cofactor engineering for advancing chemical biotechnology. *Curr Opin Biotechnol* 24 (6):994–999
2. Matsushika A, Watanabe S, Kodaki T, Makino K, Inoue H, Murakami K, Takimura O, Sawayama S (2008) Expression of protein engineered NADP plus -dependent xylitol dehydrogenase increases ethanol production from xylose in recombinant *Saccharomyces cerevisiae*. *Appl Microbiol Biotechnol* 81(2):243–255
3. Tamakawa H, Ikushima S, Yoshida S (2011) Ethanol production from xylose by a recombinant *Candida utilis* strain expressing protein-engineered xylose reductase and xylitol dehydrogenase. *Biosci Biotechnol Biochem* 75 (10):1994–2000
4. Scrutton NS, Berry A, Perham RN (1990) Redesign of the coenzyme specificity of a dehydrogenase by protein engineering. *Nature* 343 (6253):38–43
5. Cahn JKB, Werlang CA, Baumschlager A, Brinkmann-Chen S, Mayo SL, Arnold FH (2016) A general tool for engineering the NAD/NADP cofactor preference of oxidoreductases. *ACS Synth Biol* 6(2):326–333
6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235–242
7. Schrodinger LLC (2010) The PyMOL molecular graphics system, version 1.3r1
8. Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modeling. *Bioinformatics* 22:195–201
9. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 10(6):845–858
10. Engqvist MKM, McIsaac RS, Dollinger P, Flytzanis NC, Abrams M, Schor S, Arnold FH (2015) Directed evolution of *Gloeobacter violaceus* rhodopsin spectral properties. *J Mol Biol* 427(1):205–220
11. Sugimoto N, Nakano S-I, Yoneyama M, Honda K-I (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res* 24(22):4501–4505
12. Gibson DG (2011) Chapter 15: Enzymatic assembly of overlapping DNA fragments. *Methods Enzymol*, 498:349–361

Bacterial Genome Editing Strategy for Control of Transcription and Protein Stability

Ida Lauritsen, Virginia Martínez, Carlotta Ronda,
Alex Toftgaard Nielsen, and Morten H.H. Nørholm

Abstract

In molecular biology and cell factory engineering, tools that enable control of protein production and stability are highly important. Here, we describe protocols for tagging genes in *Escherichia coli* allowing for inducible degradation and transcriptional control of any soluble protein of interest. The underlying molecular biology is based on the two cross-kingdom tools CRISPRi and the N-end rule for protein degradation. Genome editing is performed with the CRMAGE technology and randomization of the translational initiation region minimizes the polar effects of tag insertion. The approach has previously been applied for targeting proteins originating from essential operon-located genes and has potential to serve as a universal synthetic biology tool.

Key words Genome editing, N-End rule pathway, CRISPR interference, CRISPR-Cas9, CRMAGE, Essential genes, N-Degron, Protein stability, PROTi, CRiPi

1 Introduction

Control over cellular protein levels is a key aspect of basic molecular biology and applied biotechnology. Nucleotide and protein-manipulation technologies are thus extremely important, best exemplified with the recent paradigm shifting CRISPR-Cas9 genome editing. Such technologies are valuable for controlling protein levels of enzymes involved in biosynthetic pathways or to elucidate the physiological functions of essential genes in which genetic knockouts are not an option [1]. Protein abundance can be manipulated by specific peptide degradation signals or “tags” that determine the protein half-life [2]. Several studies and technologies utilize synthetic degradation tags for inducible protein control engineered at the C-terminal end of proteins of interest [3–5]. However, these approaches can be suboptimal for operon-located genes (due to “polar effects” on the expression of downstream genes) and for proteins requiring a free C-terminus for

functionality. Alternative degradation signals, named N-degrons, are attached at the N-termini of proteins where stability is mostly dictated by the identity of the first amino acid and a few additional parameters [6]. N-degron residues that lead to protein degradation are called destabilizing and in bacteria, these are divided into two classes: the primary (leucine, phenylalanine, and tryptophan) and the secondary (arginine and lysine) [7]. The primary N-degrons directly promote protein degradation by adapter ClpS recognition, thereby targeting the N-degron to the ClpAP protease complex. Secondary N-degrons are modified into primary ones by enzymatic addition of the corresponding amino acids [8]. The N-end rule pathway has been identified in bacteria [7], mammals [9], yeast [6], and plants [10].

Based on the N-end rule pathway, we have recently developed and characterized a bacterial protein degradation technology, the Protein interference system (PROTi). This technology offers conditional protein degradation, involving exposure of an N-degron upon rhamnose-controlled expression of a TEV protease [11]. The N-degron is part of a small degradation tag, the PROTi tag, and the corresponding nucleotide sequence is genomically inserted in the 5'-end of genes of interest. The PROTi tag (54 nucleotides) encodes the TEV recognition site (consensus sequence: ENLYFQ↓X [12], where X denotes any amino acid except proline), the N-degron phenylalanine (in the X-position of the TEV recognition site sequence) and a linker region, shown to promote degradation by the N-end rule pathway [13] (Fig. 1). With this system, we have been able to deplete, e.g., GFP and essential proteins in *Escherichia coli* (*E. coli*)—the latter leading to a significant decrease in viability upon regrowth of colonies on agar plates after PROTi induction [11]. Combining the PROTi system with transcriptional repression by CRISPR interference (CRISPRi), we developed the CRiPi system [11]. The method of CRISPR interference (CRISPRi) enables transcriptional inhibition by use of a dead Cas9 (dCas9), assumed to prevent RNA polymerase binding [14]. By targeting the dCas9 to the DNA encoding the PROTi tag, the CRiPi system enables simultaneous protein degradation and knockdown of transcription of genes with the genomically inserted PROTi tag (Fig. 1). With this technology, we successfully knocked down soluble proteins encoded by operon-located essential genes and demonstrated enhanced phenotypes for two essential genes by simultaneous targeting of both transcription and protein stability [11].

Genomic insertion of the PROTi tag is performed using CRMAGE genome editing. This technology combines λ -Red recombineering-based multiplex-automated genome engineering (MAGE) with selection against the unmodified sequence using the CRISPR-Cas9 system in *E. coli* [15]. The small PROTi tag is encoded in a single-stranded oligonucleotide (ssDNA oligo), compatible with CRMAGE. To apply the CRISPR-Cas9 system

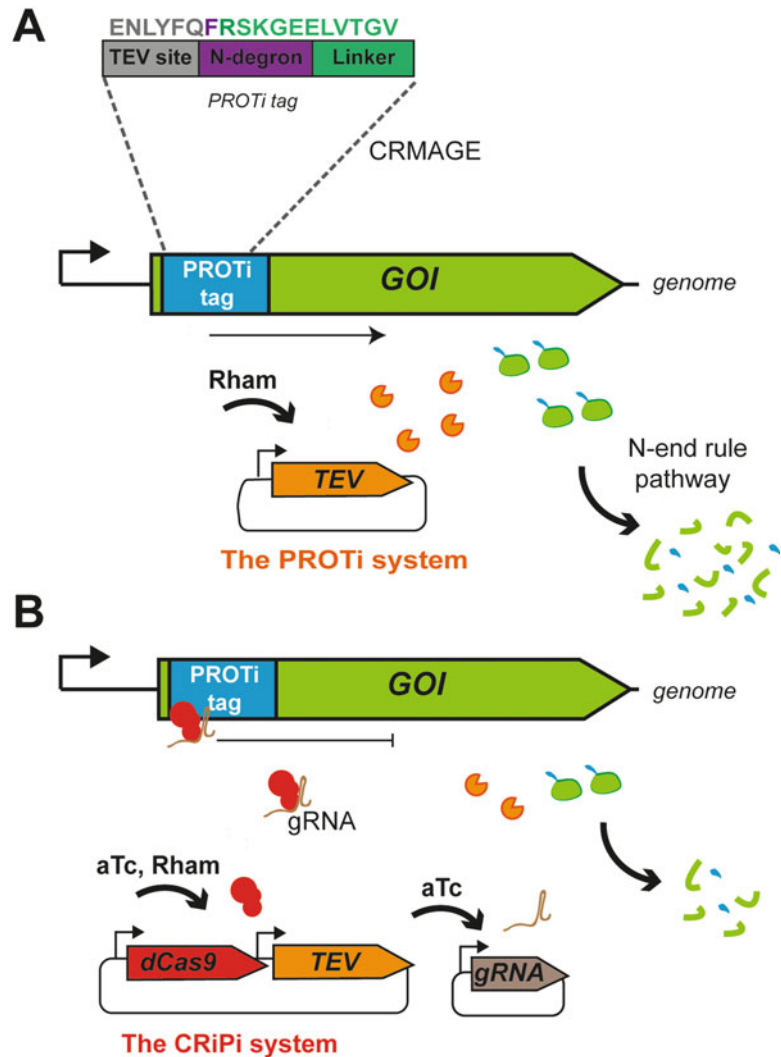


Fig. 1 The PROTi and CRiPi systems. **(a)** The PROTi tag (blue) is genomically integrated using CRMAGE [15], tagging the gene of interest (GOI) at the 5'-end. In the PROTi system, upon rhamnose-inducible expression of the TEV protease (corresponding gene and protein illustrated in orange color) the N-degron becomes exposed leading to protein degradation by N-end rule pathway (gene and protein of interest illustrated in green color). **(b)** Protein depletion can be accelerated by simultaneously implementing the CRISPRi technology. The CRiPi system relies on a combination of controlled protein degradation and repression of transcription. Expression of dCas9 (illustrated in red) can be induced by addition of anhydrotetracycline (aTc). The dCas9 is targeted towards the DNA encoding the PROTi tag by a guide RNA (gRNA)

for selection, a protospacer adjacent motif (PAM) sequence of the canonical form 5'-NGG-3' is to be removed upon recombining the ssDNA into the genome [15]. Using this selection method, identification of clones with the genomically inserted PROTi tag is

strongly facilitated. Occasionally, the genomic insertion of the PROTi tag can have lethal consequences, e.g., when targeting essential genes located in operons [11]. In that case, insertion of the PROTi tag is facilitated by simultaneous randomization of the translational initiation region (TIR), which has been shown to heavily influence translation rates [16–18]. Specifically, six nucleotides upstream and downstream from the start codon are randomized in our approach and this creates expression libraries that can be screened phenotypically [11]. The approach of TIR randomization has previously enabled tagging of seven operon-located essential genes that were not tagged before, presumably by minimizing polar effects and allowing for identification of clones with near-wild type expression levels [11]. An overview of the workflow for application of the PROTi and CRiPi systems is presented in Fig. 2.

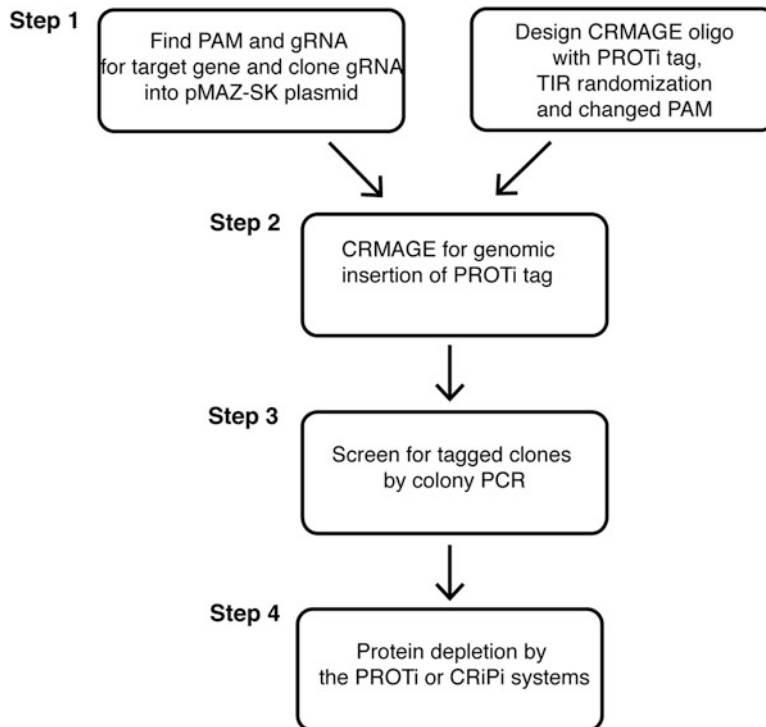


Fig. 2 Workflow for application of the PROTi and CRiPi systems. *Step 1:* Selection of a PAM and gRNA sequence in the coding sequence of the gene of interest. The selected gRNA is cloned into the pMAZ-SK plasmid. Upon expression from this plasmid, the gRNA will guide Cas9 site-specific cleavage, selecting against the unmodified sequence in CRMAGE. Furthermore, a ssDNA CRMAGE oligonucleotide is designed to encode the PROTi tag, a randomized TIR (to minimize effects of tagging) and a changed PAM to circumvent Cas9 cleavage (see Fig. 3 for further details). *Step 2:* The CRMAGE oligonucleotide and the pMAZ-SK-gRNA are used for genomic integration and selection of an inserted PROTi tag in the gene of interest with CRMAGE. *Step 3:* Genomically PROTi-tagged clones are identified by colony PCR upon recovery after CRMAGE. *Step 4:* Inducible depletion of target protein is now possible with the PROTi or CRiPi systems induced with rhamnose and/or anhydrotetracycline

2 Materials

2.1 CRMAGE

Medium, CRiPi

Inducers, and Buffers

1. LB-lennox: 10 g/L tryptone, 5 g/L yeast extract and 5 g/L NaCl supplemented with 0.5 mM MgSO₄ (Sigma-Aldrich).
2. CRiPi inducers: L-rhamnose (0.5 M), L-arabinose (10%) and anhydrotetracycline (200 µg/mL) (Sigma Aldrich).
3. NEBuffer4 (New England Biolabs® Inc.).
4. Antibiotics: Kanamycin (50 mg/mL), chloramphenicol (25 mg/mL), and ampicillin (100 mg/mL).

2.2 Oligonucleotides

1. ssDNA CRMAGE oligos (longer than 100 nucleotides) for PROTi tag insertion.
2. Oligos containing the chosen gRNA for the target gene with overhangs that match the pMAZ-SK vector [15].
3. Screening oligos: For amplification of a PCR product with, e.g., the size of 200 bp for verification of inserted PROTi tag.

2.3 Strains, Plasmids

and Polymerases

1. CRMAGE strain: *E. coli* K-12 MG1655 harboring pMA7CR.2.0 and pZS4Int-tetR [15] (available from Addgene), standard *E. coli* cloning strain (e.g., DH5α).
2. Plasmids: pMAZ-SK for gRNA cloning [15] (available from Addgene), pPROTi, pCRiPi, and pgRNA for CRiPi-induced protein depletion [11]. An overview of plasmids needed is shown in Table 1.
3. Polymerase chain reaction (PCR) kit for colony PCR (e.g., OneTaq®), Uracil-Specific Excision Reagent (USER) enzyme for cloning (New England Biolabs® Inc.).

Table 1

Plasmids for generation of PROTi-tagged genes and PROTi and/or CRiPi-induced protein depletion

| Plasmid name | Description | Reference |
|--------------|--|-----------|
| pMAZ-SK | Expression of target gRNA, aTc-inducible, Km ^R | [15] |
| pPROTi | Expression of TEV protease, rhamnose-inducible, Cm ^R | [11] |
| pCRiPi | Expression of TEV protease and dCas9, rhamnose, and aTc-inducible, Cm ^R | [11] |
| pgRNA | Expression of gRNA towards PROTi tag, aTc-inducible, Amp ^R | [11] |

3 Methods

An outline of the workflow is shown in Fig. 2.

3.1 Cloning of Target gRNA into pMAZ-SK Backbone

1. A PAM sequence in the form of 5'-NGG-3' or 5'-CCN-3' is chosen in the coding sequence, in close proximity to the start codon of the target gene (*see Note 1*). The 20 nucleotides next to the PAM sequence are the guide RNA (gRNA). If the PAM is chosen as 5'-NGG-3', the gRNA is upstream of the PAM and if selected as 5'-CCN-3', the gRNA is located downstream. Both the PAM and the gRNA are essential components for Cas9 recognition and cleavage of the target sequence.
2. Overhang sequences matching the pMAZ-SK backbone (Table 2) are added to the 20 nucleotides of the chosen gRNA (5'-GAGCAC-N₂₀-GTTTTAGAGCTAGAAAT-3') and the complementary sequence (5'-CTAAAAC-N₂₀-GTGCTCAGTATCTCT-3') according to [15] and ordered as a forward and a reverse oligo.
3. 10 µL of both the forward and reverse gRNA oligo are mixed at a concentration of 100 µM with 10 µL NEBuffer 4 and 70 µL MiliQ water in a 1.5 mL Eppendorf tube.
4. The sample is incubated at 95 °C for 5 min, then cooled to room temperature (20–25 °C) for annealing.
5. The pMAZ-SK backbone is amplified by PCR with backbone primers (Table 2) and the annealed gRNA is cloned into the pMAZ-SK plasmid by USER cloning as described in [19]. The melting temperature is 21 °C for the USER overhangs of the pMAZ-SK backbone. Cloning is performed with 15 min at 37 °C, 15 min at 21 °C and 10 min at 10 °C.
6. 5 µL of the USER reaction is mixed with 50 µL chemical competent *E. coli* cells (e.g., the DH5α strain) in a 1.5 mL

Table 2
Oligos for generation of PROTi-tagged genes and PROTi and/or CRiPi-induced protein depletion

| Oligo name | Description | Sequence (5'→3') | Reference |
|---------------------|--|--|-----------|
| gRNA overhang fw | Overhangs matching pMAZ-SK backbone after USER treatment | GAGCAC-N ₂₀ -GTTTTTA GAGCTAGAAAT | [15] |
| gRNA overhangs rv | Overhangs matching pMAZ-SK backbone after USER treatment | CTAAAAC-N ₂₀ -GTGCT CAGTATCTCT | [15] |
| pMAZ-SK backbone fw | Amplification of pMAZ-SK backbone for gRNA cloning | AGCTAGAAAUAGCAAGT TAAAATAAGGC | [15] |
| pMAZ-SK backbone rv | Amplification of pMAZ-SK backbone for gRNA cloning | AGTATCTCUATCACTGA TAGGGATGTCA | [15] |

The N₂₀ denotes the 20 nucleotides of the gRNA that is changed for the gene of interest

Eppendorf tube, put on ice for 30 min and then heat-shocked at 42 °C for 60 s.

- The sample is cooled on ice for 2 min, 500 µL of LB is added and then incubated 1 h at 37 °C with shaking.
- The sample is plated on LB agar with kanamycin (50 µg/mL) and incubated overnight at 37 °C. Positive colonies are screened by PCR followed by standard DNA purification and sequence validation.

3.2 Design of ssDNA CRMAGE Oligo with Randomized TIR and Changed PAM

An overview of CRMAGE oligo design is illustrated in Fig. 3. The CRMAGE ssDNA oligos for each gene target must be designed to target the lagging strand of the replication fork during DNA replication.

- The online program MAGE Oligonucleotide Design Tool (MODEST) is used to determine the oligo direction and chromosomal position for lagging strand-targeting. The organism *E. coli* str. K12 MG1655, version NC_000913.3 is chosen as settings [20].
- Insertion of the PROTi tag is done after the first codon downstream of the start codon of the target gene.
- To apply CRISPR-Cas9 as negative selection during CRMAGE, the wild type PAM sequence (5'-NGG-3') is changed to avoid Cas9 recognition. This substitution should only cause a synonymous substitution in the coding sequence (*see Note 2*).

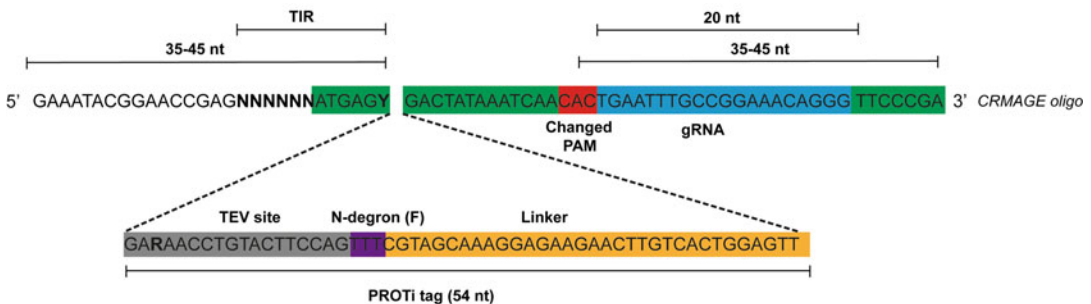


Fig. 3 Illustration of the design of a CRMAGE oligonucleotide. The CRMAGE oligonucleotide encodes the PROTi tag (illustrated in *grey*, *purple*, and *yellow*) and contains 35–45 nucleotide end homology to ensure efficient genomic insertion. The PAM sequence (*red*) is chosen in the target coding sequence (specific nucleotides shown here is a random example) and must be of the canonical form 5'-NGG-3' or 5'-CCN-3'. To avoid Cas9 recognition of clones with inserted PROTi tag, a single-nucleotide substitution is performed in one of the guanine nucleotides (or cytosine) of the PAM (shown here as the second C substituted with A). This substitution should only be a synonymous codon substitution. The translation initiation region (TIR) is completely randomized upstream of the start codon (ATG) (shown here as NNNNNN in *bold*). In addition, the first codon downstream of the start codon and the first codon of the TEV recognition site (highlighted in *bold*) are synonymously randomized (Y—cytosine or thymine in the second codon and R—adenine or guanine in the first codon of the TEV sequence). The total length of the ssDNA CRMAGE oligo typically varies from 120 to 200 nucleotides, dependent of the location of the chosen PAM sequence and gRNA. The genomic sequence of the gene *ileS* is shown here as an example

4. On each side of the chromosomal modifications (insertion of PROTi tag and change of PAM), homology regions are needed with the minimum size of 35 nucleotides (*see Note 3*).
5. For randomization of the TIR region, the six nucleotides upstream of the start codon are completely randomized. The downstream six nucleotides are randomized only to allow synonymous substitutions in the coding sequence. Due to the position for PROTi tag insertion, the first codon of the TEV-recognition site (first part of PROTi tag) is also randomized as part of the TIR region (Fig. 3).

3.3 Generation of PROTi-Tagged Strains

A modified protocol for CRMAGE is used [15]. Cultures are grown at 37 °C in a water bath shaking at 250 rpm.

1. The starting strain is *E. coli* K-12 MG1655 harboring pMA7CR.2.0 and pZS4Int-tetR plasmids [15].
2. An overnight culture is diluted to an OD₆₀₀ of 0.05 in 15 mL LB-lennox with 100 µg/mL ampicillin and 34 µg/mL chloramphenicol to select for pMA7CR.2.0 and pZS4Int-tetR.
3. When the culture reaches an OD₆₀₀ of 0.4–0.5, L-arabinose is added to a final concentration of 0.2% for expression of the λ-Red beta proteins and growth is continued for 15 min.
4. The culture is cooled on ice for 15 min, moved to 50 mL Falcon tubes, harvested by centrifugation at 6500 × *g* for 7 min at 4 °C and washed with 35 mL ice-cold MilliQ water. The harvesting and washing step is repeated.
5. After discarding the supernatant, the culture is resuspended in 1 mL ice-cold MilliQ water and transferred to a 2 mL Eppendorf tube for an additional washing step. The culture is spun at 11,000 × *g* for 1 min at 4 °C in a table-top centrifuge.
6. The sedimented cells are finally resuspended in 0.4 mL ice-cold MilliQ water.
7. For each CRMAGE reaction, 50 µL of prepared cells, 1 µL of the target ssDNA CRMAGE oligo (5 pmol/µL) and 250 ng of the corresponding pMAZ-SK plasmid with inserted gRNA are mixed in PCR tubes. The samples are electroporated at 1.8 kV, 200 Ω, 25 µF for 5 ms immediately after electroporation. Then 950 µL of LB-lennox containing 100 µg/mL ampicillin and 34 µg/mL chloramphenicol are added for recovery and the sample transferred to a new 50 mL Falcon tubes.
8. After 1 h of recovery, kanamycin is added to a final concentration of 50 µg/mL and incubated for a further 2 h.
9. Anhydrotetracycline is added to a final concentration of 200 ng/mL and incubated overnight.

10. Cultures are plated on selective LB agar plates and identification of clones with the inserted PROTi tag is performed by colony PCR with screening oligos (Table 2).

3.4 Protein Depletion by PROTi and CRiPi

1. Clones with identified PROTi tags are transformed according to standard transformation methods [21] with the plasmid-borne PROTi or CRiPi systems (Table 1).
2. From overnight cultures, 10 μ L of cells are inoculated in 1 mL LB supplied with the appropriate antibiotics in a 96-well plate and incubate at 37 °C shaking at 300 rpm.
3. L-rhamnose is added to a final concentration of 5 mM for PROTi-induced protein degradation and anhydrotetracycline to a final concentration of 200 ng/mL for transcriptional inhibition. For CRiPi induction, both rhamnose and anhydrotetracycline are added (*see Note 4*).
4. Dependent on the target protein, the phenotype of protein depletion can be measured by reduction in fluorescence, western blotting, cell viability etc. (*see Note 5*).

4 Notes

1. The PAM sequence is chosen around 3–20 nucleotides downstream of the start codon in the coding sequence to avoid interference with upstream located ribosomal binding site, sites for regulatory elements or other genes in operons.
2. Substitution of the second guanine (5'-NGG-3') to an adenine (5'-NAG-3') in the PAM sequence should be avoided due to weak Cas9 recognition.
3. The position of the chosen PAM sequence and its single-nucleotide substitution to avoid Cas9 recognition influence the total size of the CRMAGE oligo. From the changed single-nucleotide substitution in the PAM and inserted PROTi tag, the CRMAGE oligo must contain a minimum of 35-nucleotide end homology to ensure efficient chromosomal insertion. The total size of the ssDNA CRMAGE oligo should be designed as short as possible due to possible inhibitory secondary structures (e.g., hairpins) that can decrease the efficiency of chromosomal insertion.
4. Dependent on the target gene and its encoding protein and research purpose, protein depletion by the CRiPi system can be induced at different time points, e.g., corresponding to different growth stages. Protein depletion can be achieved by inducible PROTi-induced protein degradation, CRISPRi-induced transcriptional repression or by the combined CRiPi system. In our previous work, induced protein degradation of GFP was

induced after 4 h of growth. Applying CRISPRi additionally accelerated GFP depletion. For the essential genes investigated in our previous work, PROTi-induced growth inhibition was observed with rhamnose supplemented from the beginning of growth. CRISPRi-induced transcriptional inhibition for the essential genes was observed when inducing at the beginning of growth and after 4 h of growth. Synergetic CRiPi-induced protein depletion by combining transcriptional repression with protein degradation was observed for two of the tested essential genes when inducing the CRiPR system after 4 h of growth [11].

5. CRiPi-induced knockdown phenotypes can be observed in liquid culture or by plating on solid medium. Protein depletion of essential proteins was observed as a reduction in optical density and in colony forming unit numbers when plating induced culture for visualization of single colonies.

References

1. Brockman IM, Prather KLJ (2015) Dynamic metabolic engineering: new strategies for developing responsive cell factories. *Biotechnol J* 10:1360–1369
2. Schrader EK, Harstad KG, Matouschek A (2009) Targeting proteins for degradation. *Nat Chem Biol* 5:815–822
3. McGinness KE, Baker TA, Sauer RT (2006) Engineering controllable protein degradation. *Mol Cell* 22:701–707
4. Cameron DE, Collins JJ (2014) Tunable protein degradation in bacteria. *Nat Biotechnol* 32:1276–1283
5. Taxis C, Stier G, Spadaccini R et al (2009) Efficient protein depletion by genetically controlled deprotection of a dormant N-degron. *Mol Syst Biol* 5:1–7
6. Bachmair A, Finley D, Varshavsky A (1986) In vivo half-life of a protein is a function of its amino-terminal residue. *Science* 234:179–186
7. Tobias JW, Shrader TE, Rocap G et al (1991) The N-end rule in bacteria. *Science* 254:1374–1377
8. Rivera-Rivera I, Román-Hernández G, Sauer RT et al (2014) Remodeling of a delivery complex allows ClpS-mediated degradation of N-degron substrates. *Proc Natl Acad Sci U S A* 111:E3853–E3859
9. Gonda DK, Bachmair A, Wüning I et al (1989) Universality and structure of the N-end rule. *J Biol Chem* 264:16700–16712
10. Potuschak T, Stary S, Schlögelhofer P et al (1998) PRT1 of *Arabidopsis thaliana* encodes a component of the plant N-end rule pathway. *Proc Natl Acad Sci U S A* 95:7904–7908
11. Martínez V, Lauritsen I, Wolff T et al (2017) CRISPR/Cas9-based genome editing for simultaneous interference with gene expression and protein stability. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkx797>
12. Kapust RB, Tözser J, Copeland TD et al (2002) The P1' specificity of tobacco etch virus protease. *Biochem Biophys Res Commun* 294:949–955
13. Erbse A, Schmidt R, Bornemann T et al (2006) ClpS is an essential component of the N-end rule pathway in *Escherichia coli*. *Nature* 439:753–756
14. Qi LS, Larson MH, Gilbert LA et al (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* 152:1173–1183
15. Ronda C, Pedersen LE, Sommer MOA et al (2016) CRMAGE: CRISPR optimized MAGE recombineering. *Sci Rep* 6:19452
16. Kozak M (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* 361:13–37
17. McCarthy JE, Gualerzi C (1990) Translational control of prokaryotic gene expression. *Trends Genet* 6:78–85
18. Mirzadeh K, Martínez V, Toddo S et al (2015) Enhanced protein production in *Escherichia coli* by optimization of cloning scars at the vector – coding sequence junction. *ACS Synth Biol* 4:959–965
19. Cavaleiro AM, Kim SH, Seppälä S et al (2015) Accurate DNA assembly and genome

- engineering with optimized uracil excision cloning. *ACS Synth Biol* 4:1042–1046
20. Bonde MT, Klausen MS, Anderson MV et al (2014) MODEST: a web-based design tool for oligonucleotide-mediated genome engineering and recombineering. *Nucleic Acids Res* 42: W408–W415
21. Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York

Chapter 4

An Automated Pipeline for Engineering Many-Enzyme Pathways: Computational Sequence Design, Pathway Expression-Flux Mapping, and Scalable Pathway Optimization

Sean M. Halper, Daniel P. Cetnar, and Howard M. Salis

Abstract

Engineering many-enzyme metabolic pathways suffers from the design curse of dimensionality. There are an astronomical number of synonymous DNA sequence choices, though relatively few will express an evolutionary robust, maximally productive pathway without metabolic bottlenecks. To solve this challenge, we have developed an integrated, automated computational–experimental pipeline that identifies a pathway’s optimal DNA sequence without high-throughput screening or many cycles of design-build-test. The first step applies our Operon Calculator algorithm to design a host-specific evolutionary robust bacterial operon sequence with maximally tunable enzyme expression levels. The second step applies our RBS Library Calculator algorithm to systematically vary enzyme expression levels with the smallest-sized library. After characterizing a small number of constructed pathway variants, measurements are supplied to our Pathway Map Calculator algorithm, which then parameterizes a kinetic metabolic model that ultimately predicts the pathway’s optimal enzyme expression levels and DNA sequences. Altogether, our algorithms provide the ability to efficiently map the pathway’s sequence–expression–activity space and predict DNA sequences with desired metabolic fluxes. Here, we provide a step-by-step guide to applying the Pathway Optimization Pipeline on a desired multi-enzyme pathway in a bacterial host.

Key words Metabolic engineering, Synthetic biology, Operon design, Expression optimization, Kinetic modeling, Evolutionary robustness

1 Introduction

The field of metabolic engineering has reached a critical point where predictive design automation has become an essential component towards engineering multi-enzyme pathways with maximized productivities inside a target host organism. The current state-of-the-art in multi-enzyme pathway engineering is to construct several different versions of the pathway, each using different promoters [1, 2], ribosome binding sites (RBSs) [3, 4], or other

genetic parts [5, 6] that vary the pathway's enzyme expression levels. Strains carrying different pathway variants are then selected and characterized to assess their productivity, followed by analysis to determine the next set of genetic mutations that will further improve the strain's productivity. By repeating this cycle of design-build-test, incremental improvements are achieved. However, this iterative approach faces a combinatorial crisis when attempting to engineer and optimize pathways with several enzymes.

Fundamentally, optimization of a many-enzyme pathway requires searching through an astronomical number of design choices (genetic parts, and their DNA sequences) and ultimately identifying a single genetic system sequence that maximizes the pathway's productivity in a selected host. Even with the highest throughput bio-foundry and the most elegant screen or selection, it is not economically feasible to characterize even a tiny fraction of the potential design choices. It is instructive to enumerate such choices. For example, depending on its sequence, a constitutive promoter's transcription initiation rate will vary across over a 1000-fold scale and there are over 10^{18} possible 30-nucleotide promoter sequences. The translation initiation rate of a ribosome binding site (RBS) will vary across over a 100,000-fold scale, and there are over 10^{15} possible 25-nucleotide RBS sequences. Promoters and RBSs work together to determine an enzyme's overall expression level, though there are additional sequence elements, hiding inside protein coding sequences, that confound our control over an enzyme's expression level and increase the chance that evolutionary selection will break the multi-enzyme pathway's function, for example, repetitive regions or transposon insertion sites [7, 8]. A typical 100-amino acid enzyme has about 10^{50} synonymous protein coding sequences, and only a tiny fraction of them will exclude these undesired genetic elements. Promoters, RBSs, and protein coding sequences (CDSs) are also not fully modular parts. The 5' portion of the RBS sequence affects a promoter's transcription rate by altering abortive initiation [9]. The 5' portion of a CDS affects a RBS's translation rate by participating in translation-inhibiting RNA structures [10]. The CDS also affects translation elongation rates, cotranslational protein folding, and mRNA stability [11–13]. Throughout, there are additional sequence elements that affect DNA replication and genetic stability [7, 14]. Optimizing the expression, activity, and evolutionary stability of a single enzyme ultimately requires identifying an optimal promoter, RBS, and CDS sequence from 10^{81} possible choices. Engineering an 8-enzyme pathway to maximize its flux requires selecting from a 10^{550} sequence space, each with different enzyme expression levels, metabolite levels, growth toxicities, and overall pathway productivities within a particular host. At the fundamental level, a pathway's design space is vast, and intelligent approaches are essential to efficiently search and map it.

Recent pathway optimization efforts have effectively employed a divide-and-conquer strategy to search this design space. When promoters, RBSs, and CDSs are combined together to express enzymes, the design of each part was usually performed separately and isolated from one another. Promoters and RBSs may be pre-characterized, and assumed to have the same function when combined together, although that is rarely true [15, 16]. CDSs may be designed using synonymous codon optimization, but without considering other design rules. Further, the high-dimensional, quantitative relationship between enzyme expression levels and pathway productivities is not commonly determined because of the time and costs of obtaining it. As a result, there are many engineered pathways that produce product, but not at an economically competitive productivity. Commonly, determining further steps to improve productivities is not readily apparent because of the nonmodular interactions between parts, the presence of unseen, uncontrolled rate-limiting steps in enzyme expression, and the absence of quantitative relationships between enzyme expression, activity, and overall pathway productivity.

Three key technologies now offer the opportunity to rethink the fundamentals of metabolic pathway engineering and apply a more holistic design strategy. First, the costs of commercial DNA synthesis have continued to drop, particularly for sub-3000 bp DNA fragments, and scarless DNA assembly techniques have facilitated the facile construction of any genetic system regardless of sequence. Second, the development of predictive biophysical models of gene expression has enabled the design of nonnatural sequences to control protein expression levels [10, 17]. Third, kinetic metabolic modeling has reached an advanced stage where it has become possible to parameterize predictive models relating a pathway's enzyme expression levels to its overall productivity [18–21].

By combining these technologies together, we have developed an integrated computational–experimental pipeline that carries out multilevel, multi-objective optimization to determine the optimal DNA sequence of a multi-enzyme pathway that maximizes its overall productivity, while incorporating additional design criteria to ensure its genetic stability. Our pipeline does not require high-throughput screening of strain productivities, and by using relatively few experimental measurements, it determines the pathway's sequence–expression–productivity relationship, which can be reused to optimize the pathway towards different objectives.

Overall, the pipeline has four automated steps. First, our Operon Calculator algorithm is applied to design bacterial operon sequences that provide maximum control over the expression levels of all enzymes in a pathway, while eliminating all undesired genetic elements that promote genetic instability. Second, our RBS Library Calculator algorithm is employed to design optimal RBS libraries

that maximally vary enzyme expression levels with the smallest number of RBS sequence variants. Third, after constructing the operon libraries, and characterizing randomly selected variants (~100 for a 5-enzyme pathway), the measurement data is inputted into our Pathway Map Calculator, which then parameterizes a kinetic metabolic model that predicts the pathway's multidimensional sequence-expression-flux relationship. Fourth, and finally, the Pathway Map is used to determine the pathway's optimal enzyme expression levels and our RBS Calculator algorithm is used to design the RBS sequences needed to implement those optimal enzyme expression levels. Altogether, our pipeline scales exceedingly well when engineering multi-enzyme pathways; the number of needed experimental measurements is only linear with respect to the number of enzyme-catalyzed reactions in the pathway. Below, we present a "HowTo" guide to using our pipeline to engineer and optimize multi-enzyme pathways.

2 Methods

2.1 An Integrated Computational-Experimental Pipeline for Multi-enzyme Pathway Optimization

1. The necessary inputs into the pipeline are (a) the protein coding sequences (amino acids) for each heterologous enzyme in the pathway; (b) a candidate list of enzyme-catalyzed reactions that convert substrate into product, listing both endogenous and heterologous reactions that have a significant impact on product biosynthesis; and (c) a selected bacterial host. A rate law needs to be selected for each enzyme-catalyzed reaction from a dropdown list of common rate laws, including forms of allosteric regulation. If necessary, several reaction network candidates may be proposed and the pipeline will evaluate their correctness based on the measurement data-set.
2. Apply the Operon Calculator (Subheading 2.2) to design a synthetic bacterial operon sequence that expresses enzymes within the same reaction module, for example, enzymes that catalyze linear branches of the overall pathway. Repeatedly use the Operon Calculator until all enzymes in the pathway are encoded and expressed in operons. Inducible or environmentally responsive promoters may be selected during the design process to coordinate the expression of each operon.
3. Apply the RBS Library Calculator (Subheading 2.3) to design optimized RBS libraries for each enzyme CDS in the pathway. Results from the Operon Calculator are needed to carry out RBS library optimization, including the optimized RBS sequence and CDS sequences. Insert each optimized RBS library in front of its respective CDS, replacing the existing RBS sequence.

4. Append all designed operons together into a single genetic system (the “pathway”) together with desired selection markers and/or homology regions for genome integration. The genetic system will include degenerate RBS sequences, therefore specifying a library of pathway variants. Importantly, the total size of the library has been minimized, and the only variation in the library exists within the RBS sequences.
5. Construct the library of pathway variants inside a selected bacterial host. Several approaches could be used to construct such a library, for example, combining DNA synthesis of fragments, scarless fragment assembly (e.g. Gibson assembly [22], Golden Gate cloning [23], and/or MoClo cloning [24], λ_{RED} recombination [25] and/or Cas9-directed recombination [26], and/or MAGE mutagenesis [27].
6. Select a limited number of strains carrying pathway variants and measure the amount of final product (titer) produced. Use fermentation conditions that yield the most reproducible measurements while using feedstock that is the most similar to the scale-up feedstock. Calculate strain productivities (rate) by dividing titer measurements by fermentation times. Importantly, titers and productivities may be assayed using any quantitative technique that features a proportional unit scale, for example, LC/MS, GC-MS, an enzyme assay, or a fluorescent biosensor.
7. Sequence the RBSs of the characterized pathway variants, and determine their corresponding translation rates by using the RBS Library Calculator’s previous results. These translation rates are assumed to be proportional to the enzymes’ expression levels. Alternatively, proteomic or enzyme activity assays may be performed to measure the enzyme expression levels and activities.
8. Apply the Pathway Map Calculator (Subheading 2.4) to quantitatively map the pathway’s multidimensional relationship between its enzyme expression levels and end-product productivity, and to predict the enzyme expression levels that will maximize the pathway’s productivity. To do this, the Pathway Map Calculator requires a candidate reaction network (from **step 1**), the characterized pathway variants’ measured productivities (from **step 6**), and the characterized pathway variants’ predicted/measured enzyme expression levels (from **step 7**). Importantly, the Pathway Map Calculator algorithm is unit-agnostic; the solution does not depend on the techniques used to measure productivities and predict/measure enzyme expression levels so long as the same set of techniques are used throughout. Alternative candidate reaction networks may be tested for correctness, and the one with the least fit error may be used to calculate predictions.

9. Apply the RBS Library Calculator (Subheading 2.3) to implement the optimal enzyme expression levels as predicted by the Pathway Map Calculator (from **step 8**). For each enzyme CDS in the pathway, design a new RBS library (4–12 variants) that has minimum and maximum translation rates that encompass the predicted optimal expression level for the enzyme. If the curvature of the Pathway Map at the global maxima is very steep, then the RBS library should be designed to have a narrow minimum and maximum translation rate range. If the curvature is instead very broad at the global maxima, then the RBS library’s translation rates should be correspondingly broad.
10. Repeat **step 5** to incorporate the new RBS libraries upstream of their respective CDSs. Repeat **step 6** to characterize the productivities of a smaller number of pathway variants. These pathway variants are predicted to have optimal enzyme expression levels and maximal productivities. However, if there is significant deviation between productivity measurements and Pathway Map predictions, then **steps 8–10** may be repeated to further improve the Pathway Map and implement its predictions.

2.2 The Operon Calculator

The Operon Calculator automates the design of synthetic bacterial operon sequences to ensure that transcription initiation and translation initiation are the only rate-limiting steps in protein expression, while eliminating the presence of undesired genetic elements that either confound or control the operon’s protein expression levels or promote its genetic instability [13, 28]. The algorithm uses multi-objective genetic algorithm optimization with non-dominated selection to design a population of operon sequence solutions with equally optimal design criteria. The algorithm then returns the five best operon sequence solutions that satisfy the users’ design inputs, rules, and constraints.

1. Use your web browser to navigate to the Operon Calculator interface at https://salislab.net/software/OperonCalculator_ForwardDesign, and login (Fig. 1).
2. Enter a name for the operon being designed into the “Design Job/Operon Name” field.
3. Select the design rules to prioritize during your run by selecting or deselecting checkboxes under the “Active Design Rules” heading. We recommend selecting all design rules for best results. The design rules include:
 - (a) “Improve mRNA Stability”: RNase E can bind to long single-stranded RNA, while RNase III can bind to long double-stranded RNA, lowering an mRNA’s stability. This design rule identifies potential RNase binding sites,

Operon Calculator

rational design of bacterial operons to control protein expression

Design Job Title/Operon Name

Active Design Rules

| | |
|---|---|
| Improve mRNA Stability <input checked="" type="checkbox"/> | Remove Undesired Translation Products <input checked="" type="checkbox"/> |
| Remove Internal Promoters <input checked="" type="checkbox"/> | Remove Ribosomal Pause Sites <input checked="" type="checkbox"/> |
| Remove Internal Terminators <input checked="" type="checkbox"/> | Remove Genetic Instability Sites <input checked="" type="checkbox"/> |
| Remove Repetitive DNA <input checked="" type="checkbox"/> | |

Synonymous Codon Usage Table

Select a Promoter

Protein Coding Sequence #1

CDS Name:

Target Translation Rate:

CDS Sequence:

Protein Coding Sequence #2

CDS Name:

Target Translation Rate:

CDS Sequence:

Select a
 Transcriptional Terminator

Select an Organism

▶

Fig. 1 The user interface for the Operon Calculator. This interface is used in **steps 3–10**

- and activating it will minimize the presence of these sites, increasing the mRNA's stability.
- (b) “Remove Internal Promoters”: Promoter-like regions anywhere inside the operon can recruit RNA polymerase, resulting in competition for RNA polymerases and the production of erroneous mRNA transcripts. Uncontrolled erroneous transcription confounds our ability to rationally control enzyme expression levels. This design rule detects promoter-like sequences, and activating it will minimize the presence of these sequences within the operon.
 - (c) “Remove Internal Terminators”: Terminator-like sequences can prematurely terminate transcription, including both rho-independent and dependent terminators, resulting in incomplete and incorrect mRNAs. Uncontrolled erroneous transcriptional termination confounds our ability to rationally control enzyme expression levels. This design rule detects terminator-like sequences, and activating it will prevent the premature termination of transcription.

- (d) “Remove Repetitive DNA”: Long, repetitive sequences in a genetic system increase the rate of homologous recombination, whereby the intervening DNA between repeats can be excised, resulting in broken operons and pathways. This design rule identifies the presence of direct or inverted repeats, and activating it will eliminate their presence and increase the operon’s genetic stability.
 - (e) “Remove Undesired Translation Products”: Non-canonical or internal start codons can result in the translation of truncated or nonsense proteins. This design rule uses a predictive biophysical model of translation initiation to identify when an internal start codon will have a sufficiently high translation rate to produce truncated or nonsense protein. Activating this design rule will eliminate highly translated internal start codons by mutating the CDS’s synonymous codon usage.
 - (f) “Remove Ribosomal Pause Sites”: Shine-Dalgarno like sequences inside a protein coding sequence can potentially pause a ribosome during translation elongation, resulting in lower mRNA translation rates and additional sequestration of ribosomes. This design rule identifies the presence of CDS sequences that pause ribosomes, and activating it will eliminate ribosomal pause sequences by mutating the CDS’s synonymous codon usage.
 - (g) “Remove Genetic Instability Sites”: Phages and transposons integrate themselves into bacterial genomes by recognizing certain DNA sites, for example, recombinase recognition sequences or transposon insertion sites. The presence of these sites increases the risk of genetic disruption. This design rule detects the presence of genetic instability sites anywhere within the operon, and removes them, improving the genetic stability of the operon.
4. Select your preference for synonymous codons (Highly Translated or Balanced) via the “Synonymous Codon Usage” dropdown menu. For most cases, we recommend using the “Highly Translated Synonymous Codons” table, to ensure that the codon usage of a given enzyme does not limit its overall expression in the host organism.
- (a) “Highly Translated Synonymous Codons”: optimizes the coding sequence of the enzymes in the operon to preferentially use codons found in highly translated proteins in the host organism.
 - (b) “Balanced Synonymous Codons”: optimizes the coding sequence of the enzymes in the operon to use codons found in proteins with similar translation initiation rates in the host organism as the enzyme.

5. Enter your promoter sequence into the “Promoter Sequence” field, or select a promoter from the dropdown promoter menu.
6. Specify the target translation rates for each protein coding sequence in the operon in the “Target Translation Rate” fields. Specify the names of each protein coding sequence in the “CDS Name” fields. The number of CDSs in the operon can be increased or reduced using the “Add Protein CDS” or “Remove Protein CDS” links.
7. Enter your terminator sequence into the “Terminator Sequence” field, or select a terminator from the dropdown terminator menu.
8. Select the host organism from the “Select an Organism” dropdown menu. Currently, “Highly translated” and “Balanced” synonymous codon usage tables have been precalculated for *E. coli* and *B. subtilis*, though additions to the list of organisms may be requested.
9. Verify that all fields are filled and that desired design rules are selected. Click the “Submit Job” button to begin operon design. Operon designs are submitted to a queue for parallel, multi-core jobs and processed on a first-come, first-serve basis. Computation time for a 6000 bp operon is about 12 h on 10 Intel Xeon 2.6 GHz cores. Computation times are inversely proportional to the number of compute cores utilized.
10. Once the algorithm has finished, the results can be viewed under “My Results” (<https://salislab.net/software/Results>). The five best operon designs are listed (Fig. 2). Each operon result shows the designed RBS and CDS sequences, their translation initiation rates, and locations of any undesired genetic elements, including internal start codons, internal terminators, RNase binding sites, repetitive sequences, and genetic instability sites. The operon sequence designs can be exported in GenBank format by clicking on “Export as Genbank”.
11. Notes on using the Operon Calculator. In the initial design stage, we recommend using a translation rate of 100,000 to ensure that operon sequences have the ability to support very high expression levels towards maximizing pathway productivities. However, we do not advise the synthesis and molecular cloning of operons that have very high translation rates because of the risk of host toxicity, particularly if the cloning takes place on multi-copy plasmids. Instead, in a later step of the pipeline, we show how to use the RBS Library Calculator to design RBS variants that systematically vary each CDS’s translation rate. We recommend the synthesis and cloning of operons that contains RBS library variants with medium to low translation rates

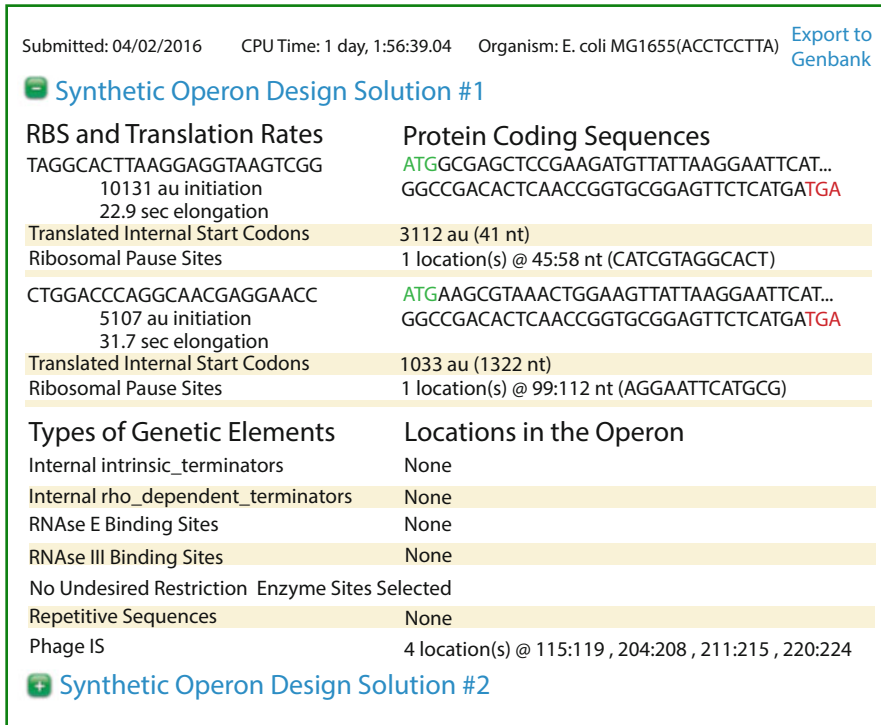


Fig. 2 The results page for the Operon Calculator. Each Synthetic Operon Design Solution shows the RBSs, optimized CDSs, and remaining genetic elements that could result in instability

(100–1000) to avoid these challenges. Where possible, we also recommend limiting the length of the operon to a maximum of 6000 nucleotides, due to increased RNA polymerase fall-off at higher transcript lengths [29]. We also recommend integration of operons into the host genome for improved genetic stability.

2.3 RBS Library Calculator

The RBS Library Calculator designs the smallest RBS library that systematically varies a CDS's translation rate across the widest possible range with a user-defined minimum and maximum translation rate [18]. We use the RBS Library Calculator to search the operon's multidimensional expression space with the smallest number of operon variants. At the design stage, before any construction, the algorithm is run for each CDS within an operon, and returns a single degenerate RBS sequence that is inserted before each CDS. The library of operon variants is then constructed through a selected synthesis, assembly, cloning, and mutagenesis strategy. Importantly, there are relatively few pinpoint nucleotide mutations between RBS variants, enabling the usage of a wider array of high-throughput mutagenesis techniques. The RBS Library Calculator also has applications beyond the pipeline described here, for example, in tuning the expression of endogenous enzymes for redirecting flux through host metabolic networks.

RBS Library Calculator_{1.0}
efficient expression optimization for multi-protein genetic systems

Title

Pre-Sequence Protein Coding Sequence

Initial RBS Sequence

Min. Translation Rate Max. Translation Rate

Target Library Size: 16 Low Res High Res

Organism

Fig. 3 The user interface for the RBS Library Calculator. This interface is used in **steps 3–7**

1. Use your web browser to navigate to the RBS Library Calculator interface at <https://salislab.net/software/RBSLibraryCalculatorSearchMode> (Fig. 3).
2. Enter a title for your RBS Library Calculator design job.
3. Using the Operon Calculator’s results, enter the nucleotide sequence of a designed CDS into the “Protein Coding Sequence” field.
4. Specify the minimum and maximum translation rates of the RBS library. We recommend selecting 100 for a minimum translation rate and at least 100,000 for a maximum translation rate. If the operons are to be genome-integrated, we recommend maximum translation rates of up to 1,000,000, though not all CDSs can support such high translation rates.
5. Select the size of the RBS library (4–64 variants) by using the “Target Library Size” slider.
6. Specify the pre-sequence that will appear before the designed RBS library. If the CDS is the first CDS in the operon, then the pre-sequence input is left blank. If the CDS is second in the operon, or further downstream, then the pre-sequence is the last 20 nucleotides of the upstream CDS that appears before the RBS library.
7. (Optional) To accelerate convergence, the algorithm allows the users to specify an RBS sequence to be used as an initial condition. We recommend utilizing the Operon Calculator’s designed RBS sequence as an initial condition for the RBS Library Calculator by copy/pasting the designed RBS

sequence into the “Initial RBS Sequence with Optional Constraints” field. There are additional ways to utilize this field.

- (a) No initial condition. If this field is left empty, the algorithm will automatically design an initial condition for optimization and no nucleotide constraints will be applied.
8. Hard constraints. Entering an RBS sequence that contains degenerate nucleotides (S/W/K/R/Y/M/D/B/H/V/N) will instruct the algorithm to only allow those nucleotides to be present at the selected positions. However, it is important to only include such hard constraints when the experimental design requires it. Otherwise, RBS libraries with limited translation rate ranges will be generated. Select the host organism from the organism menu.
9. Verify that the inputs are correct, and then click “Submit Job”. RBS library designs are submitted to a queue for parallel, multi-core jobs and processed on a first-come, first-serve basis. Computation time for a typical RBS library is about 6 h on 10 Intel Xeon 2.6 GHz cores. Computation times are inversely proportional to the number of compute cores utilized.
10. Once the algorithm has finished, the results can be viewed under “My Results” (Fig. 4). Several possible RBS library solutions are presented, shown from largest to smallest by number of RBS variants. Clicking on each RBS library solution will update the graph to show the translation initiation rates predicted for each of its members. The RBS library designs can be exported as an Excel file by clicking “Click to Export”.
11. Notes for using the RBS Library Calculator. Once the optimization begins, each iteration of the algorithm will return a check-pointed solution and the number of optimization iterations will be shown. Optimal RBS libraries are sometimes reached within a small number of iterations (10), but often-times 100 or more iterations are required. Before using results, we recommend waiting until the algorithm has completed and returned a finished solution. The RBS Library Calculator should be applied on each CDS in the operon and pathway to design customized RBS libraries controlling each enzyme’s expression level. It is NOT recommended to reuse the same RBS library for different enzyme coding sequences because changing the CDS will affect a RBS library’s translation rates.

2.4 Pathway Map Calculator

The Pathway Map Calculator creates and parameterizes kinetic metabolic models to determine a pathway’s quantitative relationship between enzyme expression levels and end-product productivity (a Pathway Map), and therefore can be used to predict the

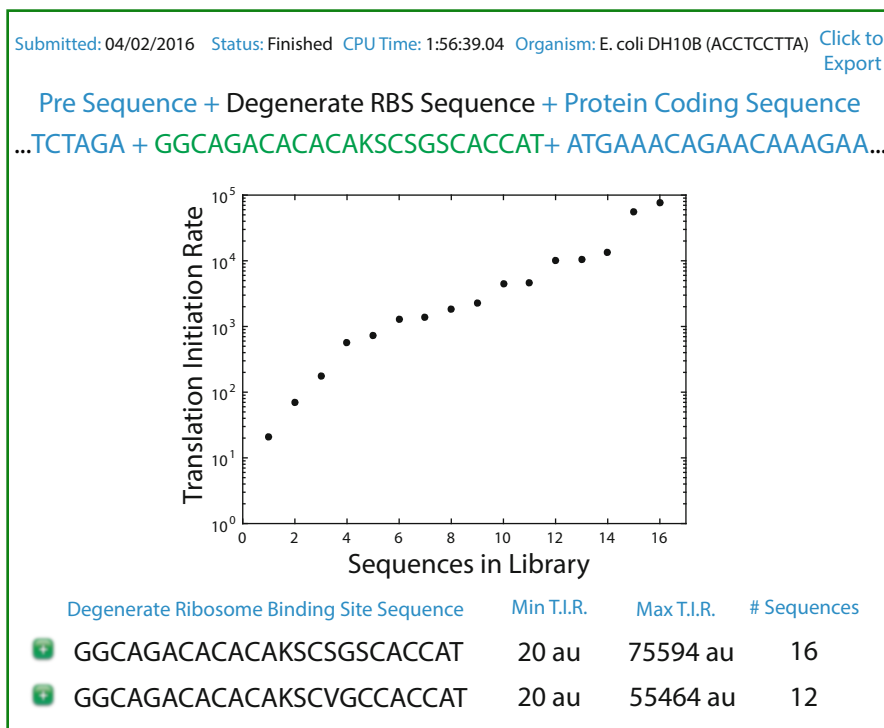


Fig. 4 The results page for the RBS Library Calculator. Each degenerate RBS library covers a wide range of Translation Initiation Rates (*center plot*), and the algorithm provides degenerate libraries of varying size to fit the user's requirements

enzyme expression levels that maximize the pathway's productivity. The algorithm only requires characterization of a small number of pathway variants because it automatically carries out several mathematical techniques to reduce the model's unknown parameters and bound their values. The number of characterized pathway variants increases only linearly with the number of reactions in the pathway; adding more enzymes to a pathway requires characterization of only a few more pathway variants. The resulting Pathway Map is a continuous, differentiable guide to optimizing the pathway's enzyme expression levels, while also quantifying the relative intrinsic kinetics of the pathway's enzymes. Therefore, the Pathway Map may also be used to sort enzymes from slowest to fastest, according to their intrinsic kinetic constants, and to prioritize the slowest enzymes for beneficial protein engineering efforts.

1. Use your web browser to navigate to the Pathway Map Calculator's interface at https://salislab.net/software/PathwayMapCalculator_CreateMode.
2. Download the Excel template by clicking the download link (Fig. 5).

A

| The Pathway Map Calculator | | | | | | | |
|----------------------------|-------------|-----------|----------|--------|-----------|---------|------|
| Name of output | Prod. units | | | | | | |
| Enzymes | Reactions | Mechanism | Subunits | Inputs | k. fluxes | Outputs | M.M. |
| | | | | | | | |
| Reaction network | | | | | | | |

B

| V. Enzymes | Ref. Path. | time | | |
|--------------------------|----------------|------|--|--|
| Enzyme expression levels | Productivities | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| User Data | | | | |

Fig. 5 Excel input spreadsheet for the Pathway Map Calculator. (a) Sheet used for inputting the stoichiometry of the pathway of interest (b) Sheet used for inputting productivity measurements and enzyme expression levels. This workbook is used in **steps 6–21**

- Enter a name for the job in the “Name of output” field. (Optional) Enter the units used for the productivity measurements in the “Productivity units” field.
- Specify the enzymes in the pathway within the “Enzymes” column.
- For each reaction catalyzed by an enzyme, enter the reactions within the “Reactions” column. Use the addition symbol (“+”) to separate substrate and product metabolites, and an arrow symbol (“->”) to separate the list of substrates from the list of products. Metabolite names should not include spaces, dashes (“-”), plus signs (“+”), or the greater than symbol (“>”). Whenever an enzyme catalyzes multiple reactions, the reactions should each be listed, and separated by a semicolon (“;”). For example, when the same enzyme catalyzes two reactions, the input designation could be: “A + B -> C; C + D -> E”.
- For each enzyme-catalyzed reaction, specify the rate law that governs the reaction’s kinetics by finding its numerical designation and placing it in the “Mechanism” column. The list of rate laws and their numerical designations can be found within the Excel template sheet named “Mechanism Directory”. If an enzyme catalyzes two reactions, then two rate law numbers must be specified, separated by a semicolon (“;”). For each

enzyme in the pathway, specify the number of subunits that bind together to form an active enzyme complex within the “Number of subunits” column. Currently, the Pathway Map Calculator only considers the number of enzyme subunits when an enzyme forms a homomeric complex. If the enzyme instead forms a heteromeric complex, then it should be approximated as a single subunit, and its designated expression level should be considered the lower of the two or more subunits.

7. Specify the precursor metabolites entering the pathway in the “Inputs” column. Specify their corresponding input flux values (if known), in molar units, within the “Known fluxes” column.
8. Specify the end-product metabolite, by-products of interest, and any other metabolite of interest within the “Outputs” column. For each metabolite of interest, specify its molar mass in the “Molar Mass” column. If the productivity measurements use molar units, then set the molar masses of all output metabolites to 1.0 to prevent improper unit conversions.
9. Switch to the “User Data” sheet of the Excel template to enter the pathway variants’ characterization data, including the list of enzymes, enzyme expression levels, and corresponding pathway variant productivities. Specify the number of enzymes in the pathway whose expression levels are being varied within the “Number of Varying Enzymes” field. On row 3, enter the names of each enzyme in the pathway and the names of the measured products. In the following rows, for each characterized pathway variant, enter the expression levels of each enzyme in a pathway alongside the measured productivities of the selected output metabolites.
10. Specify the reference pathway variant by placing its numerical designation in the “Reference Pathway” field. Ideal reference pathway variants are pathways with a highly reproducible productivity measurement (low measurement error) and moderate enzyme expression levels. The reference pathway is used to de-dimensionalize the kinetic metabolic model and provide an elementary flux constraint. It is possible to rerun the Pathway Map Calculator using different reference pathway variants to determine its effect on the Pathway Map’s predictions.
11. In the “Culture time” field, specify the fermentation/culture time to be used in the kinetic metabolic model simulation, using units of seconds.
12. Save the completed Excel file, click the “Upload Excel template” link, and select the saved Excel file. Click “Submit Job” to proceed to the next stage.
13. On the following interface, verify that all mechanistic and reaction information is correct (Fig. 6). You may also specify

| Title | | | | |
|--|----------------------|---------------------------|----------------------|----------------------|
| <input type="text"/> | | | | |
| Enzyme Name | Expression Varied? | Enzyme Catalyzed Reaction | Rate Law | Rxn dG |
| E1 | Yes | A <--> B | <input type="text"/> | <input type="text"/> |
| E2 | Yes | B + C <--> D | <input type="text"/> | <input type="text"/> |
| E3 | Yes | D <--> E | <input type="text"/> | <input type="text"/> |
| Product Name | | | | |
| Measured? | Molar Mass | | | |
| E | True | <input type="text"/> | | |
| Precursor Name | | | | |
| Known Flux | | | | |
| A | <input type="text"/> | | | |
| Model Parameterization Analysis | | | | |
| Elem Rxns + Elem Flux Modes - Measured Fluxes/Products = | Degrees of Freedom | | | |
| 9 + 1 - 1 = | 9 | | | |
| Pathway Variants = | Known Data Points | | | |
| 20 = | 20 | | | |
| <input type="button" value="Submit Job"/> | | | | |

Fig. 6 Secondary input window for the Pathway Map Calculator. Used to confirm inputs from spreadsheet and add ΔG info (if available)

the Gibbs free energy changes for each reaction in the reaction network, which will further constrain their kinetic parameter values. The interface will also report the total number of model degrees of freedom, the total number of elementary flux modes, and the number of unknown, unconstrained model degrees of freedom. To sufficiently parameterize the kinetic metabolic model, the number of characterized pathway variants should be at least two times more than the number of unknown, unconstrained model degrees of freedom.

14. Click “Submit Job” to create the Pathway Map Calculator design job. The design job is placed on a first-come, first-serve queue, shared by all multi-core, parallelized design jobs. Computation time for a typical Pathway Map Calculator run is about 12 h on 10 Intel Xeon 2.6 GHz cores. Computation times are inversely proportional to the number of compute cores utilized.
15. The results of the Pathway Map Calculator design job are found by clicking “My Results” (Fig. 7). There are three types of results. First, on the Pathway Map tab, graphs show the quantitative relationship between the expression levels of each enzyme in the pathway versus the productivities/fluxes of the output metabolites as specified in **step 8**. The optimal

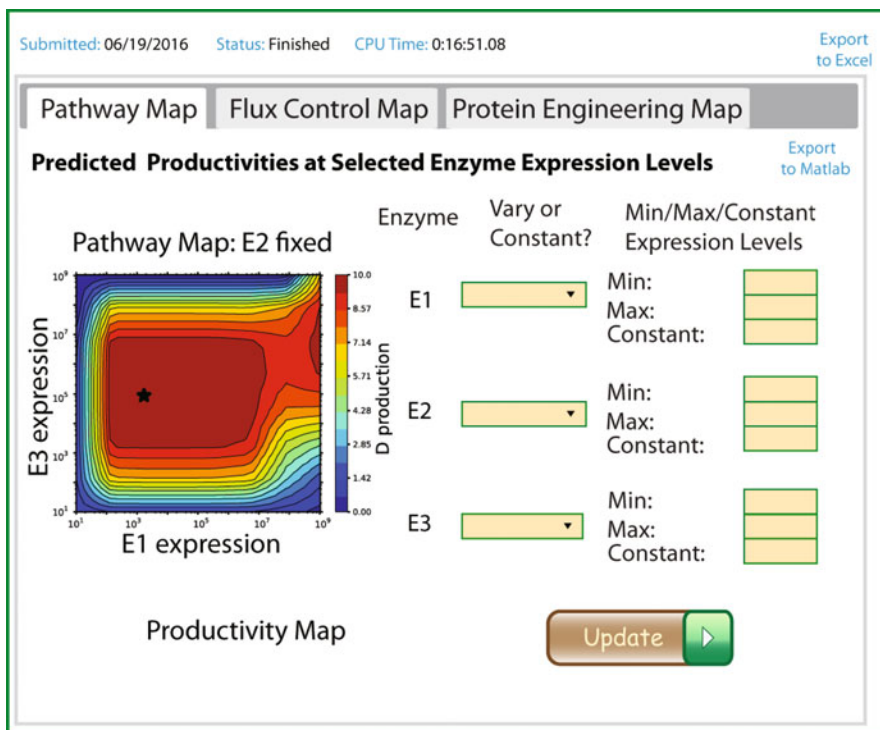


Fig. 7 The results page for the Pathway Map Calculator. Fields and buttons on the *right* are used to update the Pathway Map on the *left* with updated information. The “Flux Control Map” tab will show plots for the flux control coefficients for each enzyme, and the “Protein Engineering Map” shows which enzymes are recommended for redesign, due to flux or kinetic limitations

enzyme expression levels that maximize the pathway’s end-product productivity are designated with a star. Second, on the Flux Control Coefficient tab, graphs show how the pathway’s flux control coefficients vary, depending on each enzyme’s expression levels. Flux control coefficients (FCCs) quantify the “rate-limiting”-ness of each enzyme [30]. If an enzyme has a FCC of 0, it means that changing its expression level will have no effect on the end-product productivity. If an enzyme has an FCC of 1, it means that changing its expression level will have the most significant effect on the end-product productivity. According to metabolic control theory, the FCCs of all enzymes in a pathway must sum to 1, and therefore the rate-limitingness of each enzyme is well-quantified according to its FCC. Importantly, the enzymes’ FCCs depend on all the enzyme expression levels, and therefore the graphs show how the FCCs change when varying enzyme expression levels. Notably, the pathway will have maximal productivity when all of the enzymes have the lowest possible FCC values (0), indicating that the only rate-limiting step to improving end-product production is the precursor biosynthesis rate. Third, on the

Protein Engineering tab, the listing shows the enzymes' apparent intrinsic kinetic parameters, sorted from slowest to fastest, to provide a priority list of enzymes to be targeted for beneficial protein engineering. To properly compare enzyme kinetics across different reaction rate laws, we consider two enzymes to have equivalent intrinsic kinetics whenever they catalyze the same rate of reaction at the same substrate and enzyme concentrations and without considering allostery. However, the presence of allosteric interactions is listed here to provide protein engineering efforts with an additional target for improvement.

16. The Pathway Map Calculator's results may be downloaded as an Excel workbook via "Export to Excel" link, containing the pathway's optimal enzyme expression levels and the enzymes' intrinsic kinetic parameters. In addition, by clicking on "Export to Matlab", you may download a Matlab M-file that lists all the kinetic metabolic model's ordinary differential equations and kinetic parameters, which enables you to simulate the pathway's productivity at any selected combination of enzyme expression levels.

3 Validation and Applications

3.1 Validation of the Pathway Optimization Pipeline's Components

Many of the individual component algorithms of the Pathway Optimization Pipeline have already been applied to successfully engineer and optimize multi-enzyme pathways and gene clusters. Smanski et al. applied the RBS Library Calculator to systematically refactor and optimize a 20 enzyme nitrogen fixation gene cluster from *Klebsiella oxytoca*, increasing the rate of nitrogen fixation up to eightfold from previous efforts [31]. Lin et al. applied the RBS Library Calculator to systematically vary the expression of eight enzymes in the *E. coli* riboflavin biosynthesis pathway, increasing the final riboflavin titer by tenfold [32]. Nowroozi et al. applied the RBS Calculator and RBS Library Calculator to vary the expression of nine enzymes in the heterologous amorphadiene biosynthesis pathway in *E. coli*, increasing amorphadiene titer by fivefold [33]. Su et al. applied the RBS Library Calculator to systematically vary the expression of xylitol reductase in a heterologous xylitol biosynthesis pathway in *E. coli*, increasing the productivity of xylitol by 20-fold when hemi-cellulosic sugars were used as feedstock [34]. Ahmadi et al. used the RBS Calculator to design and optimize a 3-enzyme pathway for producing the anti-inflammatories salicylate and salicylate-2-O-B-D-glucoside at titers of 0.9 g/L and 2.5 g/L, respectively [35].

Beyond pathway engineering, the pipeline's component algorithms have also been applied to systematically optimize the

performance of a variety of sensors and genetic circuits that work together to execute complex decisions according to environmental conditions. For example, Schmidl et al. utilized the RBS Calculator to maximize the dynamic range and minimize the leakiness of bacterial light sensors, systematically varying the expression levels of the histidine sensor kinase and response regulator that control the bacteria's light sensing capability [36]. Yang et al. applied the RBS Library Calculator to systematically tune the expression of DNA recombinases (integrases) to create multi-bit memory switches, enabling the design of a genetic circuit with 1375 bytes of permanent genetic memory [37].

3.2 Validation of the Pathway Optimization Pipeline

Using our automated design pipeline, we successfully optimized the sequence the *crtEBI* pathway to improve neurosporene productivity [18]. Initially, we assembled an operon to express the *crtE*, *crtB*, and *crtI* enzymes from *R. sphaeroides* using inducible P_{lacO1} and P_{BAD} promoters and varied the expression of each enzyme using degenerate RBS libraries from the RBS Library Calculator (Fig. 8a). We characterized 73 of the resulting pathway variants for neurosporene productivity (Fig. 8c, light gray), finding that the productivity varied from 3 $\mu\text{g}/\text{DCW}/\text{h}$ to 196 $\mu\text{g}/\text{DCW}/\text{h}$. We then entered the pathway variants' predicted translation initiation rates, measured productivities, and the neurosporene biosynthesis pathway's reaction network (Fig. 8b) into the Pathway Map Calculator (Fig. 8d). The Pathway Map indicated that further increasing *crtE*, *crtB*, and *crtI* expression by specific amounts would increase the pathway's productivity by 1.5 fold, but that over-expression of all enzymes would not result in improved productivity, due to a plateau in the Pathway Map relationship. Using these insights from the Pathway Map, we applied the RBS Library Calculator to design RBS libraries that would implement translation initiation rates targeting the Pathway Map's predicted enzyme expression levels. After incorporation of these optimal RBS libraries, the pathway's productivity increased up to 286 $\mu\text{g}/\text{DCW}/\text{h}$ (Fig. 8c, dark gray).

We also examined how varying the transcription rate of the operon's promoter together with its RBS translation rates affected the pathway's overall productivity. The Pathway Map shows that changing the promoter's transcription rate will have different effects on a pathway's productivity, depending on its RBS translation rates (Fig. 8d, e). In all four characterized pathway variants, there was an optimal promoter transcription rate that maximized the pathway's productivity, though the optimal transcription rate was higher if the RBS translation rates were lower. When the pathway variants' RBS translation rates were very high (squares, diamonds), a 200-fold increase in the operons' transcription rates improved the pathways' productivities from 119 to 332 $\mu\text{g}/\text{DCW}/\text{h}$, but further promoter induction actually lowered the pathway's productivity to 254 $\mu\text{g}/\text{DCW}/\text{h}$. When the pathway variants' RBS

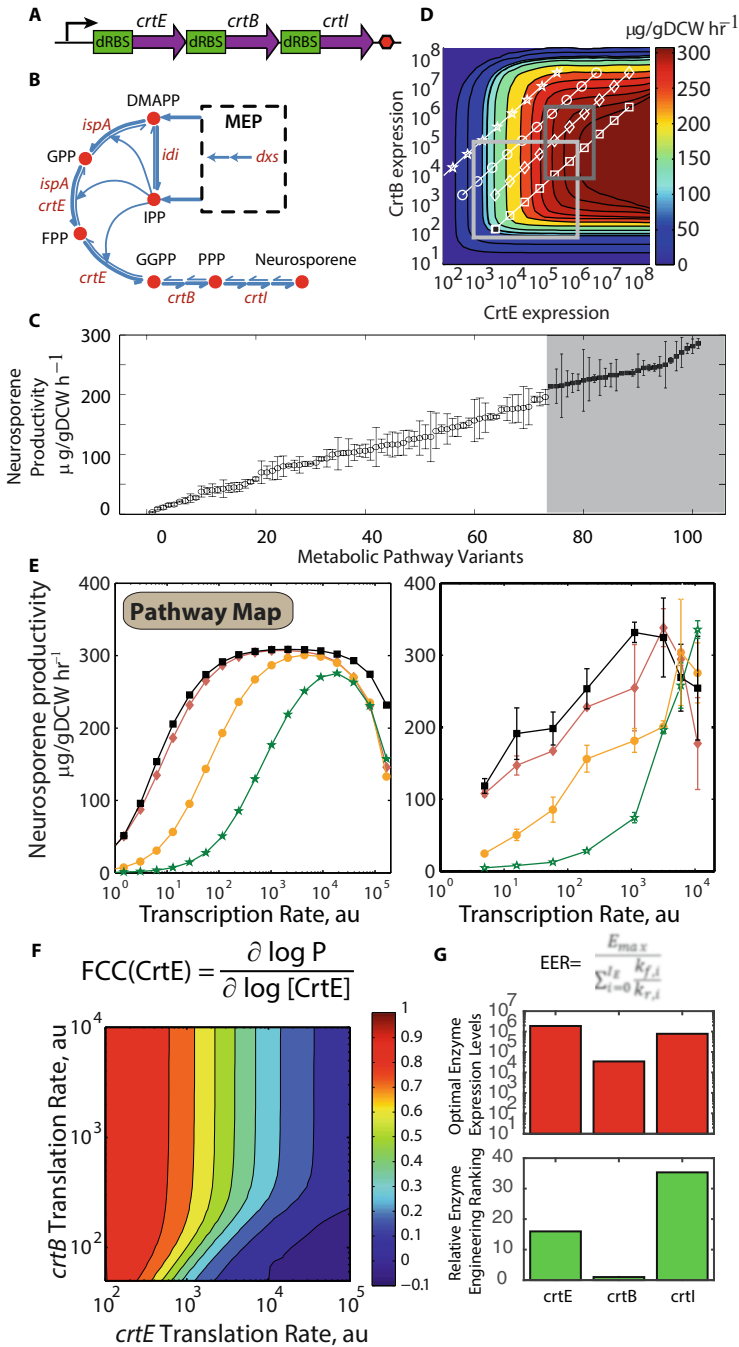


Fig. 8 The optimization of the crtEBI pathway for neurosporene production using our computational optimization approach. **(a)** The operon we designed to express neurosporene in *E. coli* via the crtEBI enzymes from *R. sphaeroides*. **(b)** The reactions performed by the crtEBI pathway used as a constraint on the resulting Pathway Map. **(c)** Pathway variants produced via degenerate RBS libraries from the RBS Library Calculator. 73 variants were initially characterized (*white*), and a subsequent optimized library was used to improve the pathway's productivity (*grey*). **(d)** Pathway Map used to optimize the expression of the crtEBI pathway. The Pathway Map

translation rates were lower (circles, stars), the promoters' transcription rates could be maximally induced without observing a plateau or decrease in productivity. Importantly, the Pathway Map was able to quantitatively predict these complex relationships between promoter transcription, RBS translation rate, and pathway productivity.

We also calculated and examined the flux control coefficients (FCCs) for the *crtEBI* pathway (Fig. 8f). As expected, the FCCs for all three enzymes were lowest when the pathway's productivity had reached its global maxima, agreeing with the Pathway Map's predictions. The FCC Map provides a quantitative approach to understanding when each enzyme becomes a more significant rate-limiting step in the overall biosynthesis pathway.

Finally, a key step to parameterizing the *crtEBI* Pathway Map is the identification of the enzymes' intrinsic kinetic parameters. By analyzing these parameters, we may then identify the enzymes that are the most intrinsically slow, and therefore would be priority targets for beneficial protein engineering efforts. Using the kinetic parameter values, we devised a quantitative metric, the Enzyme Engineering Ranking (EER), that is highest when an enzyme is intrinsically slowest (Fig. 8g). We then calculated the EER for the pathway's enzymes, and found that *crtE* and *crtI*, but not *crtB*, would be ideal targets for protein engineering. Correspondingly, the optimal enzyme expression levels for *crtE* and *crtI* are much higher than for *crtB*. Such prioritization becomes especially important when engineering many-enzyme pathways.

Altogether, after media optimization and increased *dxs* expression to improve precursor biosynthesis, the final pathway variant's productivity was 517 $\mu\text{g}/\text{gDCW}/\text{h}$.

3.3 Future Applications of the Pathway Optimization Pipeline

The Pathway Optimization Pipeline described here enables the accelerated optimization of diverse multi-enzyme pathways for broad metabolic engineering applications. Each step of the pipeline has been automated and provides actionable sequence designs and sequence improvements, freeing up researchers' time and energy to find/develop new enzymes, prototype new pathways, and create new metabolic engineering products. Importantly, the biosynthesis

Fig. 8 (continued) was fitted to 73 initial pathway variants with moderate expression levels (**d**, *light gray*), and the Pathway Map was used to inform the design of a second RBS library, spanning a more optimal expression range (**d**, *dark gray*), resulting in the improved productivities seen in **c** (*grey*). The plotted data in *white* (*square, circle, diamond, and star*) show different pathway variants at different levels of transcription. (**e**) Predicted (*left*) and measured (*right*) productivities as transcription is increased via the addition of inducer for our pathway variants (*square, circle, diamond and star*). (**f**) FCC map of the *crtEBI* pathway (**g**) Protein engineering information provided by the Pathway Map Calculator, showing the predicted optimal expression levels for each enzyme in the *crtEBI* pathway (*red*) and the relative rank of each of the enzymes for protein engineering, normalized by the lowest ranked protein, *crtB* (*green*)

pathways for many high-value natural products [38, 39] and pharmaceuticals [40, 41] are too long or complex to effectively optimize with existing approaches, resulting in nonfunctional pathways or pathways that produce too little product. The Pathway Optimization Pipeline may also be used to refactor parts of central metabolism to provide the ultimate stable, controllable chassis for expression of downstream pathways of interest [28, 42]. Furthermore, advanced pathway designs, incorporating regulatory feedback [43, 44] or biosensors [45] would be more readily achievable when using an optimally balanced pathway as the open-loop system. Our automated approach will catalyze the design and optimization of new and remarkable pathways, becoming a standard set of tools in the metabolic engineer's toolbox.

References

- Alper H et al (2005) Tuning genetic control through promoter engineering. *Proc Natl Acad Sci U S A* 102(36):12678–12683
- Mutalik VK et al (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat Methods* 10(4):354–360
- Salis HM (2011) The ribosome binding site calculator. *Methods Enzymol* 498:19–42
- Salis HM, Mirsky EA, Voigt CA (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* 27(10):946–950
- Carrier T, Jones KL, Keasling JD (1998) mRNA stability and plasmid copy number effects on gene expression from an inducible promoter system. *Biotechnol Bioeng* 59(6):666–672
- Chen Y et al (2012) Enhancing the copy number of episomal plasmids in *Saccharomyces cerevisiae* for improved protein production. *FEMS Yeast Res* 12(5):598. LP-607
- Jack BR et al (2015) Predicting the genetic stability of engineered DNA sequences with the EFM calculator. *ACS Synth Biol* 4(8):939–943
- Sleight SC et al (2010) Designing and engineering evolutionary robust genetic circuits. *J Biol Eng* 4(1):1–20
- Skancce J et al (2015) Sequence-dependent promoter escape efficiency is strongly influenced by bias for the pretranslocated state during initial transcription. *Biochemistry* 54(28):4267–4275
- Espah Borujeni A, Channarasappa AS, Salis HM (2014) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res* 42(4):2646–2659
- Espah Borujeni A, Salis HM (2016) Translation initiation is controlled by RNA folding kinetics via a ribosome drafting mechanism. *J Am Chem Soc* 138(22):7016–7023
- Grosjean H, Fiers W (1982) Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* 18(3):199–209
- Tian T, Salis HM (2015) A predictive biophysical model of translational coupling to coordinate and control protein expression in bacterial operons. *Nucleic Acids Res* 43(14):7137–7151
- Casini A et al (2014) R2oDNA designer: computational design of biologically neutral synthetic DNA sequences. *ACS Synth Biol* 3(8):525–528
- Davis JH, Rubin AJ, Sauer RT (2011) Design, construction and characterization of a set of insulated bacterial promoters. *Nucleic Acids Res* 39(3):1131–1141
- Kosuri S et al (2013) Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proc Natl Acad Sci U S A* 110(34):14024–14029
- Brewster RC, Jones DL, Phillips R (2012) Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*. *PLoS Comput Biol* 8(12):e1002811
- Farasat I et al (2014) Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria. *Mol Syst Biol* 10:731–731
- Khodayari A et al (2014) A kinetic model of *Escherichia coli* core metabolism satisfying

- multiple sets of mutant flux data. *Metab Eng* 25:50–62
20. Theisen MK, Lafontaine Rivera JG, Liao JC (2016) Stability of ensemble models predicts productivity of enzymatic systems. *PLoS Comput Biol* 12(3):e1004800
 21. Tran LM, Rizk ML, Liao JC (2008) Ensemble modeling of metabolic networks. *Biophys J* 95(12):5606–5617
 22. Gibson DG et al (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6(5):343–345
 23. Engler C, Kandzia R, Marillonnet S (2008) A one pot, one step, precision cloning method with high throughput capability. *PLoS One* 3(11):1–7
 24. Weber E et al (2011) A modular cloning system for standardized assembly of multigene constructs. *PLoS One* 6(2):e16765
 25. Murphy KC (1998) Use of bacteriophage – recombination functions to promote gene replacement in *Escherichia coli*. *J Bacteriol* 180(8):2063–2071
 26. Jiang Y et al (2015) Multigene editing in the *Escherichia coli* genome using the CRISPR-Cas9 system. *Appl Environ Microbiol* 81(7):2506–2514
 27. Wang HH et al (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460(7257):894–898
 28. Ng CY et al (2015) Rational design of a synthetic Entner-Doudoroff pathway for improved and controllable NADPH regeneration. *Metab Eng* 29:86–96
 29. Forde NR et al (2002) Using mechanical force to probe the mechanism of pausing and arrest during continuous elongation by *Escherichia coli* RNA polymerase. *Proc Natl Acad Sci U S A* 99(18):11682–11687
 30. Fell DA (1998) Increasing the flux in metabolic pathways: a metabolic control analysis perspective. *Biotechnol Bioeng* 58(2–3):121–124
 31. Smanski MJ et al (2014) Functional optimization of gene clusters by combinatorial design and assembly. *Nat Biotechnol* 32(12):1241–1249
 32. Lin Z et al (2014) Metabolic engineering of *Escherichia coli* for the production of riboflavin. *Microb Cell Factories* 13:104
 33. Nowroozi FF et al (2014) Metabolic pathway optimization using ribosome binding site variants and combinatorial gene assembly. *Appl Microbiol Biotechnol* 98(4):1567–1581
 34. Su B et al (2015) Efficient production of xylitol from hemicellulosic hydrolysate using engineered *Escherichia coli*. *Metab Eng* 31:112–122
 35. Ahmadi MK et al (2016) *E. coli* metabolic engineering for gram scale production of a plant-based anti-inflammatory agent. *Metab Eng* 38:382–388
 36. Schmidl SR et al (2014) Refactoring and optimization of light-switchable *Escherichia coli* two-component systems. *ACS Synth Biol* 3(11):820–831
 37. Yang L et al (2014) Permanent genetic memory with >1-byte capacity. *Nat Methods* 11(12):1261–1266
 38. Zhou J et al (2014) Engineering *Escherichia coli* for selective geraniol production with minimized endogenous dehydrogenation. *J Biotechnol* 169:42–50
 39. Moon TS et al (2009) Production of glucaric acid from a synthetic pathway in recombinant *Escherichia coli*. *Appl Environ Microbiol* 75(3):589–595
 40. Ajikumar PK et al (2010) Isoprenoid pathway optimization for Taxol precursor overproduction in *Escherichia coli*. *Science (New York, NY)* 330(6000):70–74
 41. Thodey K, Galanie S, Smolke CD (2014) A microbial biomanufacturing platform for natural and semisynthetic opioids. *Nat Chem Biol* 10(10):837–844
 42. Brockman IM, Prather KLJ (2015) Dynamic knockdown of *E. coli* central metabolism for redirecting fluxes of primary metabolites. *Metab Eng* 28:104–113
 43. Soma Y, Hanai T (2015) Self-induced metabolic state switching by a tunable cell density sensor for microbial isopropanol production. *Metab Eng* 30:7–15
 44. Xu P et al (2014) Improving fatty acids production by engineering dynamic pathway regulation and metabolic control. *Proc Natl Acad Sci U S A* 111(31):11299–11304
 45. Fang M et al (2016) Intermediate-sensor assisted push-pull strategy and its application in heterologous deoxyviolacein production in *Escherichia coli*. *Metab Eng* 33:41–51

Computational Approaches on Stoichiometric and Kinetic Modeling for Efficient Strain Design

Mohammad Mazharul Islam and Rajib Saha

Abstract

Engineering biological systems that are capable of overproducing products of interest is the ultimate goal of any biotechnology application. To this end, stoichiometric (or steady state) and kinetic models are increasingly becoming available for a variety of organisms including prokaryotes, eukaryotes, and microbial communities. This ever-accelerating pace of such model reconstructions has also spurred the development of optimization-based strain design techniques. This chapter highlights a number of such frameworks developed in recent years in order to generate testable hypotheses (in terms of genetic interventions), thus addressing the challenges in metabolic engineering. In particular, three major methods are covered in detail including two methods for designing strains (i.e., one stoichiometric model-based and the other by integrating kinetic information into a stoichiometric model) and one method for analyzing microbial communities.

Key words Computational strain design, Stoichiometric metabolic models, Kinetic models, Metabolic modeling, Ensemble modeling, Microbial communities, Whole plant model, Multi-tissue model

1 Introduction

Engineering carbon-efficient microbial strains for overproduction of biochemicals is the overarching goal in biotechnology spanning biofuels, biorenewables, secondary metabolites, and drugs [1]. The design of such strains with product yields as close as the theoretical maximum generally requires multiple, iterative, and, often, nonintuitive genetic interventions. A growing number of computational strain design procedures [2–11] relying on mathematical optimization framework have emerged benefiting from rapid advancements in the reconstruction of genome-scale metabolic (GSM) models [12, 13], thus addressing the challenge of identifying and quantifying these interventions and minimizing the counteractions of the organisms in response to them [1]. The large number of computational strain design approaches can be broadly categorized into two

main strategies: steady-state modeling of metabolism and kinetic modeling of metabolism.

Flux Balance Analysis (FBA) is a universally used approach for studying GSM models and subsequently applying them for metabolic engineering purposes [14–16]. For performing FBA, a GSM reconstruction needs to be represented in the form of tabulation of stoichiometric coefficients (known as stoichiometric matrix or S-matrix), in which each column represents a single metabolite whereas each row signifies a specific reaction. Under pseudo-steady state, FBA assumes that the internal concentration of metabolites within a cellular system stays constant over time [14]. In addition to the mass balance constraints, environmental constraints based on availability of nutrients, electron acceptors, or other environmental conditions, relation of reaction rates with concentrations of metabolite, and negative free energy change for spontaneous reactions can also be imposed. The effects of gene expressions may result in regulatory constraints on these models as the cell adapts to environmental changes [17]. The solution space of this under-determined system of equations represents the bounds of metabolic flux distribution that the cell can achieve under a given condition [16, 18]. An optimization-based strain design algorithm can next be used with specific objective functions (such as cellular growth rate and yield of a desired bioproduct) to simulate biological behavior of the cell.

Microbial production of a desired chemical often requires engineering of the gene content and circuitry of the microbial host in order to redirect the metabolic flow toward the product of interest, which might be in a direct competition with the cell growth. Computational strain design protocols (both stoichiometric and kinetic) aim at the system-wide identification of intervention strategies for the enhanced production of bioproducts in microorganisms [8]. The linearity of the underlying mathematics in stoichiometric GSM models and minimal requirement of biological knowledge and data significantly contribute to computational savings and tractability, even for large models with several thousands of active reactions involved. These approaches have taken a multitude of directions, ranging from knocking out some functionalities (genes) [2], forcing metabolite fluxes by using genetic manipulations to attain a pre-specified overproduction target of a biochemical [7], introducing non-native functionalities [10], optimizing carbon, energy or cost efficiency of the cellular machinery to push it to the limits governed by thermodynamics [19], flux modulations [5], minimization of metabolic adjustments [20], optimal metabolic and regulatory gene deletions as well as metabolic gene over-expressions coupling the biomass production and product formation [6], among many others. However, stoichiometric models lack the ability to capture metabolite concentration information,

enzyme saturation, and nonlinearities due to kinetic and regulatory effects [1].

Integrating metabolic fluxes, metabolite concentrations, and kinetic variables in a unified constraint-based formulation is aimed to increase the quantitative prediction capacity of FBA. Incorporation of experimental and theoretical bounds on thermodynamic and kinetic variables ensures that the predicted steady-state fluxes are thermodynamically and biochemically feasible [21]. On the other hand, kinetic models of metabolism quantify the reaction fluxes as functions of metabolite concentrations, enzyme levels, and kinetic parameters (related to enzyme turnover, saturation, and allosteric regulation) [1]. They require quantitative expressions that link reaction fluxes to metabolite concentrations. A system of ordinary differential equations is typically solved to obtain the temporal variation in metabolite concentrations and reaction fluxes. Different forms of mechanistic expressions (e.g., Michaelis-Menten or Hill Kinetic expressions) have been used extensively, which require knowledge of detailed enzyme function mechanism and characterization [22–24]. Therefore, advance in kinetic modeling efforts have been limited by available measurement techniques, availability of data, and computational expenses. However, kinetic models are advantageous in predicting metabolic behavior and dynamics of any organism or a group of organisms at any condition far from steady state. Although kinetic modeling of large-scale metabolic networks can be challenging, one revolutionary approach to address this bottleneck is to build an ensemble of dynamic models [25] that reach the same steady state over time using enzyme-expression levels.

In this chapter, we present a comprehensive summary of the stoichiometric and kinetic model-based strain design methods that have been applied to both eukaryotic and prokaryotic organisms. We also discuss about the frameworks used for multi-tissue plant metabolic modeling and microbial community modeling. In the Notes section, we provide a step-by-step guideline and mathematical formulations of three major modeling frameworks that will enable the reader to use these modeling tools.

2 Modeling of Single-Species Metabolism (*see Note 2*)

Genome-scale metabolic models summarize the known metabolic information of organisms in a mathematically defined reaction network. A key objective of utilizing these models for metabolic engineering is to improve the production of bioproducts by providing genetic intervention strategies in terms of gene knock outs/ins and up/down regulations [26]. In this section, we highlight a few of the major modeling efforts and strain engineering techniques that can be applied on individual organisms (i.e., both prokaryotes and eukaryotes). Figure 1 shows a typical

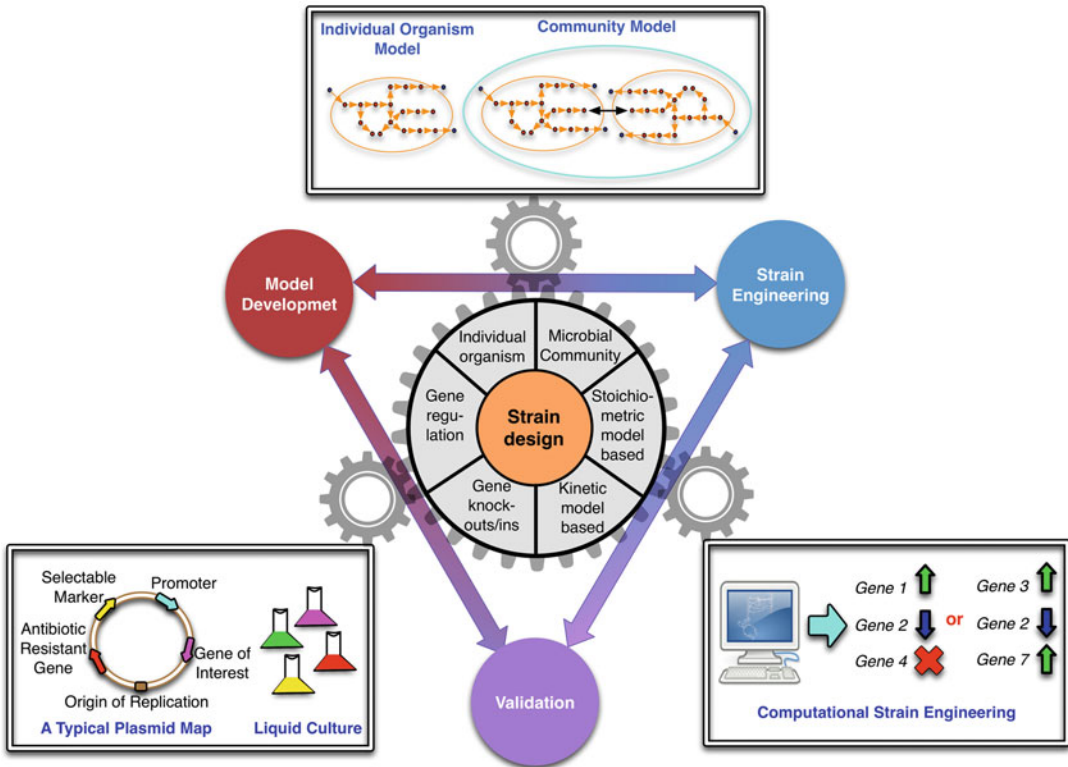


Fig. 1 A visual representation of the computational strain design workflow. In silico metabolic model(s) of individual organisms or multispecies communities are fed into computational strain engineering algorithms, upon which stoichiometric or kinetics-based genetic perturbation strategies are identified. The results in terms of gene regulation and knock-in/knockouts are then carried out in wet lab, and the discrepancies are used to refine models for the overproducing strains

computational strain design workflow by highlighting three major components: model development, strain design, and experimental validation.

2.1 Prokaryotic Models

The existing computational strain design techniques can broadly be classified into two categories based on the underlying models on which these can be applied: (a) Stoichiometric model-based strain engineering and (b) Kinetic model-based strain engineering (*see Notes 3–7*). Here, we summarize the techniques that are either developed recently or applied broadly for metabolic engineering applications.

2.1.1 Strain Design Methods Using Stoichiometric Models

Steady-state modeling of cellular metabolism and designing intervention strategies are popular for a multitude of reasons; the foremost being the strategies in practice tend to operate inside the cellular factory at a steady state or a pseudo-steady state, with a minimal reaction-level dynamics. To this end, a kinetic model and steady-state model of the same metabolic system under the same

condition would make identical predictions for reaction rates at steady state [27]. Therefore, a strain design algorithm based on steady-state modeling is preferable when answering most of the challenging questions in identifying an overproduction strain.

1. OptKnock

OptKnock [2] aims to design reaction eliminations that reshape network connectivity in such a way that the production of the target metabolite is maximized, while ensuring that a drain towards growth resources (i.e., carbon source, redox potential, and energy). It uses a bi-level optimization framework, which explicitly accounts for two competing objective functions (metabolic engineering objective as the outer objective, and cellular fitness objective as the inner objective). The outer problem in OptKnock identifies reaction candidates for elimination that maximize the production of the target metabolite (i.e., product), while the inner problem redistributes metabolic fluxes so as to maximize the biomass formation in the perturbed network subject to the reaction eliminations imposed by the outer problem. The bi-level optimization problem is nonlinear because one of the constraints is implicitly expressed as an optimization problem. Therefore, one needs to convert this bi-level structure to a standard single-level optimization problem. This can easily be achieved by aggregating the constraints of the inner problem with those of its dual while imposing the strong duality condition. OptKnock has been used to suggest gene deletion strategies for overproduction of succinate, lactate, and 1,3-propanediol in *E. coli* by Pharkya et al. [28] and other researchers [29, 30].

2. RobustKnock

OptKnock procedure selects the most optimistic value for the product yield, which might be uncoupled from growth, and therefore undesirable. To address this challenge, Tepper and Schlomi [3] introduced a modified three-level optimization framework named RobustKnock, which optimizes the worst case of the product formation while biomass production is maximized. The outer max–min problems aim to identify gene knockouts that maximize the minimum/guaranteed yield of the desired biochemical, whereas the inner problem is similar to OptKnock and maximizes the cellular objective for a set of knockouts (*see Note 8*). RobustKnock procedure has been applied to a genome-scale model of *E. coli* (*i*JR904) [31] to identify optimal knockout strategies for overproduction of 52 different chemicals.

3. OptGene

OptGene [32] is an evolutionary programming-based (i.e., Genetic Algorithm or GA) method for rapidly identifying gene

deletion strategies for maximizing a desired phenotypic objective function. It presents two major advantages: higher computational speed and ability to optimize for nonlinear objective functions. GA formulation can provide with multiple solutions, and thus an opportunity to choose from alternate solutions. The objective function value can be calculated using FBA, minimization of metabolic adjustment (MOMA) [20], regulatory on-off minimization (ROOM) [33] or any other algorithm since GA is independent of scoring algorithm used. It has been used to identify gene-deletion strategies for improving yield and substrate-specific productivity of three metabolites (i.e., vanillin, glycerol and succinate) in a genome-scale model of *S. cerevisiae* [34].

4. OptForce

OptForce [7] identifies potential genetic interventions by classifying metabolic reactions (in nonredundant sets of individual, pairs, triples, quadruples, etc.) depending on whether their flux values must change (i.e., increase, decrease, or equal to zero) to meet a pre-specified overproduction target. The sets of the identified interventions are then utilized to efficiently extract a minimal set of fluxes that must actively be forced through genetic manipulations (i.e., FORCE set) to ensure that all fluxes in the network are consistent with the overproduction objective. The efficiency and quality of engineering interventions predicted by OptForce depend on the available metabolic flux analysis (MFA [35–37]) data for the wild-type strain. OptForce framework has been employed to recommend intervention strategies in *E. coli* model *iAF1260* [38] for maximizing succinate production. OptForce framework can also be applied in the intermediate stages of strain design by recalculating the set of engineering interventions as new flux data for one or multiple mutant strains become available [7]. For example, Ranganathan et al. [39] presented an iterative metabolic engineering effort integrating OptForce predictions and MFA techniques to meet the challenge of tuning overproduction for specific free short-chain fatty acids.

2.1.2 Strain Design Methods Using Kinetic Models

Integration of kinetic information with stoichiometric modeling or developing kinetic models offers potential advantages for improved phenotype prediction and consequently more precise computational strain design [40]. Here, we summarize two such methods.

1. k-OptForce

k-OptForce [8] incorporates available kinetic descriptions for the reactions in OptForce [7] to sharpen the prediction of intervention strategies for improving the production of a chemical of interest (*see Note 9*). It enables identification of a minimal set of interventions comprised of both enzymatic parameter

changes (for reactions with available kinetics) and reaction flux changes (for reactions with only stoichiometric information). Application of k-OptForce identifies interventions that tend to cause less dramatic rearrangements of the flux distribution so as not to violate concentration bounds. k-OptForce is also capable of finding nonintuitive interventions aiming at alleviating the substrate-level inhibition of key enzymes in order to enhance the flux towards the product of interest, which cannot be captured by stoichiometry-based analysis. k-OptForce algorithm has successfully been applied to the overproduction of L-serine in *E. coli* and triacetic acid lactone (TAL) in *S. cerevisiae* [8].

2. Ensemble Modeling

Ensemble modeling [25] provides a promising remedy to address the challenges of identifying kinetic parameter values for a large number of reactions and kinetics for a wide range of environmental and genetic perturbations. The collection of ensemble allows for the examination of possible phenotypes of the network upon perturbations, such as changes in enzyme expression levels. Availability of perturbation data for different phenotypes of the same species can improve the mechanistic framework to generate more accurate predictions, and the ensemble converges to a smaller set of models. This approach is advantageous because it avoids the need for detailed characterization of kinetic parameters and generates a set of models that describes relevant phenotypes upon enzyme perturbations with acceptable accuracy. Zomorodi et al. [11] has proposed an optimization-based algorithm for systematic identification of genetic perturbations and populating the ensemble of models with relevant model parameterizations after each round of model screening. More recently, a formal parameter optimization approach is proposed by Khodayari et al. [41] that minimizes the discrepancies between model predictions and flux measurements using both wild-type and multiple mutant strains fluxomic data sets. Ensemble modeling approach has been demonstrated on *E. coli* primary metabolism [25], improved in terms of iterative parameterization [11] as well as parameter optimization [41], and integrated with k-OptForce to identify minimal sets of interventions in *E. coli* for succinate overproduction [42].

2.2 Eukaryotic Models

Compared to the number of prokaryotic models, there are not many eukaryotic models currently available. The major reasons include, but not limited to, the scarcity of knowledge regarding gene annotations, intra- and intercellular transporters, subcellular locations of genes/proteins, and complexity involved in interactions among major cell/tissue types. Except for the kinetic models of *S. cerevisiae* (budding yeast) and Chinese Hamster Ovary (CHO)

cells, these models are mostly used for characterizing metabolism or physiology, rather than for engineering. Herein, we briefly review the existing efforts in terms of modeling and engineering of eukaryotic organisms.

2.2.1 Stoichiometric Models of Eukaryotic Organisms

Community-driven consensus model building efforts exist for yeast [43]. These models are still being regularly updated and yeast metabolism is being characterized so that in future computational strain engineering techniques can be applied in yeast to generate testable hypotheses. Global plant models (by considering the entire plant as one single unit) are available for arabidopsis [44], maize [45], sorghum [46], sugarcane [46], rapeseed [47], and rice [48]. These models, in essence, analyze the physiology and/or metabolism of these plants. In addition, tissue-specific models such as models of barley seed [49] and maize leaf [50] try answering specific questions including the states of barley seed under anoxic, hypoxic, and aerobic conditions and the metabolic/physiological conditions of maize leaf under nitrogen sufficient and deficient conditions. In case of CHO model, metabolic behaviors related to the physiological changes under growth and nongrowth conditions have been analyzed [19]. Going beyond the single tissue level, Grafahrend-Belau and coworkers have developed a whole-plant model of barley to study the sink-to-source relationship shift of the stem when leaf source capacity is decreased [51].

2.2.2 Kinetic Models of Eukaryotic Organisms

Parachin and coworkers have developed a kinetic model for breaking down xylose by introducing two different xylose catabolic pathways *S. cerevisiae* and predicted increased xylulokinase activity for improving the xylose consumption [52]. The in silico result has further been experimentally validated by showing a 27–37% increase in xylose utilization. Chen and coworkers have reconstructed kinetic models of cytosolic glucose metabolism map for *S. cerevisiae* and CHO cells to propose gene knockout in order to increase the production of DHAP, which is a major intermediate for synthesizing fine chemicals [53]. In a recent work, a dynamic metabolic model for the recombinant CHO-DXB11 cell line producing the EG2-hFc monoclonal antibody has been developed to quantify extracellular substrates and metabolites concentration among others for two batch and two fed-batch cultures [54]. Another more recent work on CHO cell kinetic model applies multi-objective dynamic optimization for identifying intervention strategies in terms of enzymatic modifications, up- and downregulations to achieve increased productivity of antibody along with simultaneous improvement of titer, biomass level, and decrease of byproduct (i.e., lactate and ammonia) levels [55]. In plants, Wang and coworkers have built a small kinetic model of monoglignol biosynthesis in *Populus trichocarpa* to predict the effect of pathway enzymes on lignin content and composition and also elucidate the regulation involved in lignin biosynthesis [56].

3 Modeling of Multispecies Metabolism

Microorganisms, in nature, exist and function in diverse, robust, integrated, and interactive consortia instead of living in isolation. The interactions among the members of such a community in the form of unidirectional and/or bidirectional exchange of biochemical cues and their temporal variations (due to environmental perturbations) result in one or more population(s) being benefited from a thorough cooperation or negatively affected (e.g., competition for limiting resources). They can even have a combination of both [57–61] and, thereby, change community composition, structure, function, and stability [62–65]. Like the eukaryotic modeling efforts, there have been a growing number of metabolic modeling frameworks to understand and elucidate the inter-species interactions in simple microbial communities and their dynamics. We highlight a few such frameworks in this section.

3.1 *Steady-State Modeling of Microbial Communities*

A number of modeling frameworks based on constraint-based metabolic models have been developed recently to study the steady-state behavior of simple microbial consortia [62, 65–71]. In some of these efforts the metabolic networks of the two microorganisms are treated as separate compartments similar to eukaryotic metabolic models, and the exchange of metabolites are accommodated using a separate compartment [65–67, 72, 73]. Graph-theoretic approach and the minimization of metabolic adjustment (MOMA) hypothesis [20] have also been used to computationally identify co-growth and mutualistic relationships among community members [62, 68, 71]. Nagarajan et al. [69] have proposed a multi-omics modeling workflow combining genomic, transcriptomic, and physiological data with GSM models to assess inter-species electron transfer in a syntrophic microbial community comprising of two *Geobacter* species. A number of other approaches including elementary mode analysis, evolutionary game theory, nonlinear dynamics, and stochastic processes [74–80] have been attempted to model such communities. All these methods are based on optimization problems with a single objective function, and therefore, are not able to capture the multilevel nature of decision making in microbial communities. To capture the trade-offs between species-level fitness and community level objective in microbial communities, Zomorodi and Maranas have introduced OptCom [57] with a multilevel and multi-objective optimization formulation.

OptCom is a multilevel and multi-objective optimization framework that postulates a separate biomass maximization problem for each community member as the inner problems. The inter-species interactions are modeled by using inter-organism flow constraints in the outer problem imposing a flux balance constraint in the extracellular environment for each metabolite shared by the

community members. The objective function of the outer problem represents a community-level fitness criterion (e.g., maximization of community biomass), or surrogates a desired bioengineering objective (*see* **Note 10**). OptCom can include any type of interactions (positive, negative or combinations thereof) and also accommodate any number of microbial species (or guilds) in its framework. As a case study, OptCom has been applied to quantify the syntrophic associations in two-species microbial systems, assess the optimality level of growth in phototrophic microbial mats, and elucidate the inter-species metabolite and electron transfer in a synthetic microbial community [57].

3.2 Dynamic Modeling of Microbial Communities

Microbial communities change with time and also in response to environmental or artificial perturbations [81]. The development of dynamic community-level models is a challenging task due to the increased complexity and incomplete knowledge about the dynamics of interspecies interactions over a changing environment. There have been a number of recent efforts aimed to address this challenge [68, 71, 82–86], most of which are based on the extension of dynamic flux balance analysis (dFBA) for single species [87–89]. Dynamic Multispecies Metabolic Modeling (DMMM) proposed by Zhuang et al. [83] has been used to model the competition between *Rhodospirillum rubrum* and *Geobacter sulfurreducens* in an anoxic subsurface environment, and to devise a long-term practical bioremediation strategy for groundwater uranium reduction [82]. Minty et al. [63] have introduced a new modeling strategy in which the dynamic co-growth of two species is simulated by solving a series of batch culture problems over time and hierarchical clustering techniques are used to reveal feasible region of substrate partition between competing community members.

d-Optcom [90], introduced by Zomorodi et al., incorporates the dynamic mass balance equations and substrate uptake kinetics in the OptCom modeling framework [57] and enables the direct assessment of the shared metabolites and biomass concentrations in a given community. d-OptCom procedure has been used to capture the dynamics of the co-growth of *E. coli* auxotrophic mutant pairs, elucidate the impact of the addition of a new member to a uranium-reducing microbial community on its growth, dynamics and composition, and to assess the efficacy of lactate and acetate injection to enhance uranium reduction.

4 Notes

This section provides necessary steps regarding how anyone will decide on the modeling strategies and strain engineering approaches to follow. At the beginning of this section, we define sets, parameters, and variables that are common across all strain

design algorithms. We next present a set-by-step decision-making process and explain a few representative techniques.

1. General definition of Sets, Parameters, and Variables:

Sets

J = Set of reactions.

I = Set of metabolites.

Parameters

S_{ij} = Stoichiometric coefficient of metabolite $i \in I$ in reaction $j \in J$.

LB_j and UB_j = Lower and upper bound on reaction $j \in J$.

r_i^{uptake} = Defined rate of uptake of the metabolite i .

r_i^{export} = Defined rate of export of the metabolite i .

K = Number of knockouts.

Variables

v_j = Flux of reaction j .

v_i^{uptake} = Uptake flux of the metabolite i .

v_i^{export} = Export flux of the metabolite i .

y_j = Binary variable associated with knocked-out and active reactions.

$$y_j = \begin{cases} 0, & \text{reaction is knocked out} \\ 1, & \text{reaction is active.} \end{cases}$$

c_i = Concentration of metabolite i .

2. Although genome-scale metabolic models can be developed for different types of organisms or biological systems (i.e., individual/multi-tissue/multispecies), strain design techniques, thus far, have been developed and subsequently applied mostly for individual organisms.
3. First step in selecting strain design and optimization procedure is to make a decision on whether a steady-state or dynamic analysis is to be performed. For steady-state strain design algorithms for single species, go to **step 4**. For kinetic approaches to strain design, go to **step 5**.
4. If a deterministic strain design algorithm is preferred, choose from steady-state strain design algorithms based on gene deletion (OptKnock or RobustKnock), addition (OptStrain), and over- or under-expression (OptForce). If Regulatory information is available, use OptORF. If a heuristic method is preferred, use OptGene. Here, we describe RobustKnock in **step 8**.
5. If kinetic information about reactions in the network is available, one may opt to use the more restrictive kinetic strain design approach like k-OptForce or a detailed parameterization procedure of Ensemble modeling. We describe k-OptForce in **step 9**.

6. For steady-state analysis of multispecies communities, a compartmentalized approach is sufficient for some communities, for example, in case of pure competition. Otherwise, to elucidate the multitude of interactions among community members and understand the trade-offs, use OptCom (as explained in **step 10**).
7. For dynamic analysis of multispecies communities, if a compartmentalized approach is sufficient use DMMM, otherwise use d-OptCom,.
8. RobustKnock addresses the limitation of OptKnock in identifying growth-coupled product maximization by optimizing the worst-case scenario of the product formation while biomass production is maximized. It is a bi-level max–min optimization that searches for a set of gene knockouts under which the minimal production rate of a target chemical is maximized while satisfying the inner level FBA solution space. The bi-level max–min optimization problem of RobustKnock is shown below.

$$\begin{array}{ll} \text{Maximize} & \text{Minimize} \\ y \in \{0,1\}^q & \nu_j \end{array} \quad \nu_{\text{Chemical}}$$

$$i^T y \geq m - k$$

subject to

$$\left[\begin{array}{l} \text{Maximize} \quad \nu_j \quad \nu_{\text{biomass}} \\ \text{subject to} \\ \sum_{j \in J} S_{ij} \cdot \nu_j = 0 \quad \forall i \in I \quad (1) \\ \nu_{\text{biomass}} \geq \nu_{\text{biomass threshold}} \\ \text{LB}_j \leq \nu_j \leq \text{UB}_j \quad \forall j \in I \quad (2) \\ \nu_{\text{uptake}, i} = r_{\text{uptake}, i} \quad \forall i \in I_{\text{uptake}} \quad (3) \\ \nu_{\text{export}, i} = r_{\text{export}, i} \quad \forall i \in I_{\text{export}} \quad (4) \end{array} \right]$$

Where, q is the number of reactions in the network before the reversible reactions are split into two separate irreversible reactions, and m is the number of irreversible reactions following the split. The outer max–min problem identifies for a set of k number of gene knockouts size that maximizes the minimal (guaranteed) production rate of the target chemical (ν_{chemical}). The standard FBA formulation in the inner problem searches for a feasible solution space (flux distribution) with maximal biomass production rate, given a set of knocked-out reactions from the outer max–min problem. The set of reactions that are knocked-out are denoted by the binary variables $y \in \{0, 1\}^q$

($y_i = 0$ if reaction i is knocked-out, and 1 if not). A lower bound on the biomass production rate based on a pre-computed threshold, denoted by $v_{\text{biomass_threshold}}$, is used to obtain a biologically feasible flux distribution that enables growth. This threshold is set to a fraction (10%, 20% or 30% of the maximal possible biomass production rate) under the given growth medium. The bi-level max–min problem is solved via a two-step procedure: (1) transforming the bi-level max–min problem into a max–min problem (similar to that used in OptKnock); (2) transforming the resulting max–min problem into a mixed-integer linear programming (MILP) problem. The solution of this optimization problem identifies the candidates for gene knockouts for which the binary variable y equals to 0. These reactions can now be removed from the organism to develop the overproducing strain.

9. The k-OptForce algorithm builds upon the OptForce framework and integrates available kinetic information. The genome-scale metabolic network is partitioned into two subsets: reactions with kinetic information $J^{\text{kin}} = \{j \mid j = 1, 2, \dots, N^{\text{K}}\}$, and reactions linked only by stoichiometry $J^{\text{stoic}} = \{j \mid j = 1, 2, \dots, N^{\text{S}}\}$. Portion of the metabolic network with kinetic information is mathematically described by a system of nonlinear ODEs indicating the mass balance for each metabolite, by solving which one can obtain steady-state metabolite concentration and fluxes. The feasible space of the remaining portion of the network J^{stoic} is constructed to be consistent with the predicted fluxes of the reactions in J^{kin} . The system of ODEs is solved first to obtain steady-state fluxes for reactions in J^{kin} . The phenotypic space of the reference strain is then identified by iteratively maximizing and minimizing the flux of each reaction in J^{stoic} , while keeping the fluxes of reactions in J^{kin} fixed at their steady-state values, and restricting the flux of any other reaction for which any experimental data is available at their experimentally determined values or ranges.

$$\begin{array}{ll}
 \text{minimize/maximize} & v_j & \forall j \in J \\
 \text{subject to} & \sum_j S_{i,j} v_j = 0 & \forall i \in I \\
 & v_j^{\text{data}, L} \leq v_j \leq v_j^{\text{data}, U} & \forall j \in J^{\text{data}} \\
 & v_j = v_j^{\text{uptake}} & \forall j \in J^{\text{uptake}} \\
 & v_j = v_j^{\text{SS}} & \forall j \in J^{\text{kin}} \\
 & \text{LB}_j \leq v_j \leq \text{UB}_j & \forall j \in J \\
 & v_{\text{biomass}} = v_{\text{biomass}}^{\text{max, ref}} &
 \end{array}$$

The phenotypic space of the overproducing strain consistent with a desired overproduction target is similarly constructed by successively maximizing and minimizing the flux of network reactions subject to network stoichiometry, overproduction target, while also incorporating the kinetic expressions for reactions in J^{kin} as additional constraints. A v_j^{max} value below $v_j^{\text{max,ref}}$ indicates downregulation, while an increase of the same indicates upregulation of the enzyme-catalyzed reaction.

$$\begin{array}{lll}
\text{minimize/maximize} & v_j & \forall j \in J \\
\text{subject to} & \sum_j S_{i,j} v_j = 0 & \forall i \in I \\
& v_j \geq v_i^{\text{target}} & \forall j \in J^{\text{targets}} \\
& v_{\text{biomass}}^{\text{min}} \leq v_{\text{biomass}} & \\
& v_j = v_j^{\text{uptake}} & \forall j \in J^{\text{uptake}} \\
& \text{LB}_j \leq v_j \leq \text{UB}_j & \forall j \in J \\
& v_j = v(v_j^{\text{max}}, c_i, k_j^p) & \forall j \in J^{\text{kin}}, i \in I^{\text{kin}}, p \in P \\
& c_i^{\text{min}} \leq c_i \leq c_i^{\text{max}} & \forall i \in I^{\text{kin}} \\
& 0 \leq v_i^{\text{max}} \leq z \cdot v_i^{\text{max,ref}} & \forall j \in J^{\text{kin}}
\end{array}$$

The sets of reactions that must be upregulated ($MUST^U$), downregulated ($MUST^L$), or be knocked out ($MUST^X$) are identified by contrasting the flux space of the wild-type network with that of the overproducing strain. Higher order MUST sets are identified through contrasting the sums and difference of two or more fluxes (depending on the order) between the reference and the desired phenotypes.

The FORCE sets are identified by solving a bi-level optimization problem similar to the original OptForce formulation. The outer problem maximizes the flux towards the desired chemical consistent with reaction kinetics and stoichiometry. Binary variables $y^{L,\text{kin}}$ and $y^{U,\text{kin}}$ associated with the $MUST^L$ and $MUST^U$ sets of reactions in J^{kin} respectively, are used to propagate the effect of engineering modifications to the corresponding v_j^{max} . If $y^{L,\text{kin}}$ equals 1, v_j^{max} for that reaction can be downregulated to a value between 0 and its wild-type $v_j^{\text{max,ref}}$. If $y^{U,\text{kin}}$ equals 1, v_j^{max} for that reaction is upregulated to a value between $v_j^{\text{max,ref}}$ and a z -fold change in it. Otherwise, v_j^{max} is kept unaffected at their reference $v_j^{\text{max,ref}}$ value. At the same time, the inner problem simulates the worst-case scenario minimizing the product formation. A separate set of binary

variables $y^{L,\text{stoic}}$ and $y^{U,\text{stoic}}$ identify interventions in J^{stoic} is required to guarantee a nonzero yield of the target chemical consistent with the flux distribution in J^{kin} . It is important to note that while the fluxes in J^{kin} impart the flux distribution for the fluxes in J^{stoic} , they remain unaffected by the worst-case simulation of the reactions in J^{stoic} in the inner problem.

Interventions predicted by k-OptForce (along with alternative manipulation strategies) need to be carefully scrutinized and manually curated to understand the underlying metabolic reason (complex nonlinear couplings between distal reactions through metabolite pools) behind each intervention. The following checks can be useful in general: (1) check whether any metabolite participating in affected reactions is hitting lower or upper concentration bounds, (2) identify if a common metabolite is fixing the branching ratio of fluxes in two pathways, (3) resolve whether a metabolite is limiting the flux of a reaction through substrate-level inhibition, (4) confirm if the flux in a pathway has been restricted because the upper bound of one of the reactions has hit its upper bound, and, (5) analyze all alternate intervention strategies to identify common metabolites and/or enzymes that are being targeted [8].

10. The mathematical description of the OptCom procedure [57] is given in below.

Maximize/Minimize $z =$ Community-level objective

subject to

$$\left[\begin{array}{l} \text{Maximize} \\ v_j^k \\ \text{subject to} \\ \sum_{j \in J^k} S_{ij}^k \cdot v_j^k = 0 \quad \forall i \in \mathbf{I}^k \quad (1) \\ \text{LB}_j^k \leq v_j^k \leq \text{UB}_j^k \quad \forall j \in \mathbf{I}^k \quad (2) \\ v_{\text{uptake},i}^k = r_{\text{uptake},i}^k \quad \forall i \in \mathbf{I}_{\text{uptake}}^k \quad (3) \\ v_{\text{export},i}^k = r_{\text{export},i}^k \quad \forall i \in \mathbf{I}_{\text{export}}^k \quad (4) \end{array} \right] \quad \forall k \in \mathbf{K}$$

$$\sum_k r_{\text{uptake},i}^k + e_i^c = \sum_k r_{\text{export},i}^k + u_i^c \quad \forall i \in \mathbf{I}_{\text{shared}} \quad (5)$$

$$r_{\text{uptake},i}^k, r_{\text{export},i}^k, e_i^c, u_i^c \geq 0 \quad \forall i \in \mathbf{I}_{\text{shared}}, k \in \mathbf{K}$$

The inner problem(s) represents the steady-state flux balance problem for each microorganism (or guild) k with limits on uptake or export flux of a shared metabolite at the values $r_{\text{uptake},i}^k$ and $r_{\text{export},i}^k$, respectively, which are imposed by the outer problem. Constraint (5) in the outer problem establishes a mass

balance for each shared metabolite present in the extracellular environment (shared metabolite pool), where the terms $\sum_k r_{\text{uptake},i}^k$ and $\sum_k r_{\text{export},i}^k$ represent the total uptake and export of the shared metabolite i by community members, respectively. This constraint is the key equation for modeling the interactions and communications among participants of the community. OptCom formulation can be modified to identify the minimal number of direct interventions (i.e., knock-up/down/outs) to the community members which maximizes the production of a desired compound (e.g., by considering the overproduction of desired compound as the outer problem objective function), thus extending the applicability of strain design tools like OptKnock and OptForce. towards a community-oriented metabolic engineering protocol.

Acknowledgement

This work was supported by the University of Nebraska-Lincoln faculty start-up grant 21-1106-4308 to R.S.

References

1. Maranas CD, Zomorodi AR (2016) Optimization methods in metabolic networks. Wiley, NJ
2. Burgard AP, Pharkya P, Maranas CD (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng* 84(6):647–657. doi:[10.1002/bit.10803](https://doi.org/10.1002/bit.10803)
3. Tepper N, Shlomi T (2010) Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. *Bioinformatics* 26(4):536–543. doi:[10.1093/bioinformatics/btp704](https://doi.org/10.1093/bioinformatics/btp704)
4. Kim J, Reed JL, Maravelias CT (2011) Large-scale bi-level strain design approaches and mixed-integer programming solution techniques. *PLoS One* 6(9):e24162. doi:[10.1371/journal.pone.0024162](https://doi.org/10.1371/journal.pone.0024162)
5. Pharkya P, Maranas CD (2006) An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metab Eng* 8(1):1–13. doi:[10.1016/j.ymben.2005.08.003](https://doi.org/10.1016/j.ymben.2005.08.003)
6. Kim J, Reed JL (2010) OptORF: optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. *BMC Syst Biol* 4:53. doi:[10.1186/1752-0509-4-53](https://doi.org/10.1186/1752-0509-4-53)
7. Ranganathan S, Suthers PF, Maranas CD (2010) OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Comput Biol* 6(4):e1000744. doi:[10.1371/journal.pcbi.1000744](https://doi.org/10.1371/journal.pcbi.1000744)
8. Chowdhury A, Zomorodi AR, Maranas CD (2014) k-OptForce: integrating kinetics with flux balance analysis for strain design. *PLoS Comput Biol* 10(2):e1003487. doi:[10.1371/journal.pcbi.1003487](https://doi.org/10.1371/journal.pcbi.1003487)
9. Cotten C, Reed JL (2013) Constraint-based strain design using continuous modifications (CosMos) of flux bounds finds new strategies for metabolic engineering. *Biotechnol J* 8(5):595–604. doi:[10.1002/biot.201200316](https://doi.org/10.1002/biot.201200316)
10. Pharkya P, Burgard AP, Maranas CD (2004) OptStrain: a computational framework for redesign of microbial production systems. *Genome Res* 14(11):2367–2376. doi:[10.1101/gr.2872004](https://doi.org/10.1101/gr.2872004)
11. Zomorodi AR, Lafontaine Rivera JG, Liao JC, Maranas CD (2013) Optimization-driven identification of genetic perturbations accelerates the convergence of model parameters in ensemble modeling of metabolic networks. *Biotechnol J* 8(9):1090–1104. doi:[10.1002/biot.201200270](https://doi.org/10.1002/biot.201200270)

12. Palsson B (2006) *Systems biology : properties of reconstructed networks*. Cambridge University Press, Cambridge; New York
13. Oberhardt MA, Palsson BØ, Papin JA (2009) Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* 5:320–320. doi:10.1038/msb.2009.77
14. Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? *Nat Biotechnol* 28(3):245–248. <http://www.nature.com/nbt/journal/v28/n3/abs/nbt.1614.html>—supplementary-information
15. Oberhardt MA, Chavali AK, Papin JA (2009) Flux balance analysis: interrogating genome-scale metabolic networks. *Methods Mol Biol* 500:61–80. doi:10.1007/978-1-59745-525-1_3
16. Varma A, Palsson BO (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl Environ Microbiol* 60(10):3724–3731
17. Terzer M, Maynard ND, Covert MW, Stelling J (2009) Genome-scale metabolic networks. *Wiley Interdiscip Rev Syst Biol Med* 1(3):285–297. doi:10.1002/wsbm.37
18. Varma A, Palsson BO (1993) Metabolic capabilities of *Escherichia coli*: I. Synthesis of biosynthetic precursors and cofactors. *J Theor Biol* 165(4):477–502. doi:10.1006/jtbi.1993.1202
19. Chowdhury A, Maranas CD (2015) Designing overall stoichiometric conversions and intervening metabolic reactions. *Sci Rep* 5:16009. doi:10.1038/srep16009
20. Segrè D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A* 99(23):15112–15117. doi:10.1073/pnas.232349399
21. Fleming RM, Thiele I, Provan G, Nasheuer HP (2010) Integrated stoichiometric, thermodynamic and kinetic modelling of steady state metabolism. *J Theor Biol* 264(3):683–692. doi:10.1016/j.jtbi.2010.02.044
22. Kadir TA, Mannan AA, Kierzek AM, McFadden J, Shimizu K (2010) Modeling and simulation of the main metabolism in *Escherichia coli* and its several single-gene knockout mutants with experimental verification. *Microb Cell Factories* 9:88. doi:10.1186/1475-2859-9-88
23. Rizzi M, Baltés M, Theobald U, Reuss M (1997) In vivo analysis of metabolic dynamics in *Saccharomyces cerevisiae*: II. Mathematical model. *Biotechnol Bioeng* 55(4):592–608. doi:10.1002/(SICI)1097-0290(19970820)55:4<592::AID-BIT2>3.0.CO;2-C
24. Heinrich R, Rapoport TA (1974) A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *Eur J Biochem* 42(1):89–95
25. Tran LM, Rizk ML, Liao JC (2008) Ensemble modeling of metabolic networks. *Biophys J* 95(12):5606–5617. doi:10.1529/biophysj.108.135442
26. Zomorodi AR, Suthers PF, Ranganathan S, Maranas CD (2012) Mathematical optimization applications in metabolic networks. *Metab Eng* 14(6):672–686. doi:10.1016/j.ymben.2012.09.005
27. Zielinski DC, Palsson BØ (2012) Kinetic modeling of metabolic networks. In: Wittmann C, Lee YS (eds) *Systems metabolic engineering*. Springer Netherlands, Dordrecht, pp 25–55. doi:10.1007/978-94-007-4534-6_2
28. Pharkya P, Burgard AP, Maranas CD (2003) Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock. *Biotechnol Bioeng* 84(7):887–899. doi:10.1002/bit.10857
29. Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, Maranas CD, Palsson BO (2005) In silico design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol Bioeng* 91(5):643–648. doi:10.1002/bit.20542
30. Yim H, Haselbeck R, Niu W, Pujol-Baxley C, Burgard A, Boldt J, Khandurina J, Trawick JD, Osterhout RE, Stephen R, Estadilla J, Teisan S, Schreyer HB, Andrae S, Yang TH, Lee SY, Burk MJ, Van Dien S (2011) Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nat Chem Biol* 7(7):445–452. doi:10.1038/nchembio.580
31. Reed JL, Vo TD, Schilling CH, Palsson BO (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol* 4(9):R54–R54
32. Patil KR, Rocha I, Forster J, Nielsen J (2005) Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics* 6:308. doi:10.1186/1471-2105-6-308
33. Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A* 102(21):7695–7700. doi:10.1073/pnas.0406346102
34. Förster J, Famili I, Fu P, Palsson BØ, Nielsen J (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res* 13(2):244–253. doi:10.1101/gr.234503
35. Zamboni N, Fendt SM, Ruhl M, Sauer U (2009) (13)C-based metabolic flux analysis.

- Nat Protoc 4(6):878–892. doi:[10.1038/nprot.2009.58](https://doi.org/10.1038/nprot.2009.58)
36. Dalman T, Wiechert W, Noh K (2016) A scientific workflow framework for (13)C metabolic flux analysis. *J Biotechnol* 232:12–24. doi:[10.1016/j.jbiotec.2015.12.032](https://doi.org/10.1016/j.jbiotec.2015.12.032)
 37. Wiechert W, Mollney M, Petersen S, de Graaf AA (2001) A universal framework for 13C metabolic flux analysis. *Metab Eng* 3(3):265–283. doi:[10.1006/mben.2001.0188](https://doi.org/10.1006/mben.2001.0188)
 38. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 3:121. doi:[10.1038/msb4100155](https://doi.org/10.1038/msb4100155)
 39. Ranganathan S, Tee TW, Chowdhury A, Zomorodi AR, Yoon JM, Fu Y, Shanks JV, Maranas CD (2012) An integrated computational and experimental study for overproducing fatty acids in *Escherichia coli*. *Metab Eng* 14(6):687–704. doi:[10.1016/j.ymben.2012.08.008](https://doi.org/10.1016/j.ymben.2012.08.008)
 40. Chowdhury A, Khodayari A, Maranas CD (2015) Improving prediction fidelity of cellular metabolism with kinetic descriptions. *Curr Opin Biotechnol* 36:57–64. doi:[10.1016/j.copbio.2015.08.011](https://doi.org/10.1016/j.copbio.2015.08.011)
 41. Khodayari A, Zomorodi AR, Liao JC, Maranas CD (2014) A kinetic model of *Escherichia coli* core metabolism satisfying multiple sets of mutant flux data. *Metab Eng* 25:50–62. doi:[10.1016/j.ymben.2014.05.014](https://doi.org/10.1016/j.ymben.2014.05.014)
 42. Khodayari A, Chowdhury A, Maranas CD (2015) Succinate overproduction: a case study of computational strain design using a comprehensive *Escherichia coli* kinetic model. *Front Bioeng Biotechnol* 2. doi:[10.3389/fbioe.2014.00076](https://doi.org/10.3389/fbioe.2014.00076)
 43. Heavner BD, Smallbone K, Price ND, Walker LP (2013) Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. *Database (Oxford)* 2013:bat059. doi:[10.1093/database/bat059](https://doi.org/10.1093/database/bat059)
 44. de Oliveira Dal’Molin CG, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK (2010) AraGEM, a genome-scale reconstruction of the primary metabolic network in *Arabidopsis*. *Plant Physiol* 152(2):579–589. doi:[10.1104/pp.109.148817](https://doi.org/10.1104/pp.109.148817)
 45. Saha R, Suthers PF, Maranas CD (2011) *Zea mays* iRS1563: a comprehensive genome-scale metabolic reconstruction of maize metabolism. *PLoS One* 6(7):e21784. doi:[10.1371/journal.pone.0021784](https://doi.org/10.1371/journal.pone.0021784)
 46. Dal’Molin CG, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK (2010) C4GEM, a genome-scale metabolic model to study C4 plant metabolism. *Plant Physiol* 154(4):1871–1885. doi:[10.1104/pp.110.166488](https://doi.org/10.1104/pp.110.166488)
 47. Pilalis E, Chatziioannou A, Thomasset B, Kolisis F (2011) An in silico compartmentalized metabolic model of *Brassica Napus* enables the systemic study of regulatory aspects of plant central metabolism. *Biotechnol Bioeng* 108(7):1673–1682. doi:[10.1002/bit.23107](https://doi.org/10.1002/bit.23107)
 48. Poolman MG, Kundu S, Shaw R, Fell DA (2013) Responses to light intensity in a genome-scale model of rice metabolism. *Plant Physiol* 162(2):1060–1072. doi:[10.1104/pp.113.216762](https://doi.org/10.1104/pp.113.216762)
 49. Grafahrend-Belau E, Schreiber F, Koschutski D, Junker BH (2009) Flux balance analysis of barley seeds: a computational approach to study systemic properties of central metabolism. *Plant Physiol* 149(1):585–598. doi:[10.1104/pp.108.129635](https://doi.org/10.1104/pp.108.129635)
 50. Simons M, Saha R, Amiour N, Kumar A, Guillard L, Clement G, Miquel M, Li Z, Mouille G, Lea PJ, Hirel B, Maranas CD (2014) Assessing the metabolic impact of nitrogen availability using a compartmentalized maize leaf genome-scale model. *Plant Physiol* 166(3):1659–1674. doi:[10.1104/pp.114.245787](https://doi.org/10.1104/pp.114.245787)
 51. Grafahrend-Belau E, Junker A, Eschenroder A, Muller J, Schreiber F, Junker BH (2013) Multiscale metabolic modeling: dynamic flux balance analysis on a whole-plant scale. *Plant Physiol* 163(2):637–647. doi:[10.1104/pp.113.224006](https://doi.org/10.1104/pp.113.224006)
 52. Parachin NS, Bergdahl B, van Niel EW, Gorwa-Grauslund MF (2011) Kinetic modelling reveals current limitations in the production of ethanol from xylose by recombinant *Saccharomyces cerevisiae*. *Metab Eng* 13(5):508–517. doi:[10.1016/j.ymben.2011.05.005](https://doi.org/10.1016/j.ymben.2011.05.005)
 53. Chen N, Koumpouras GC, Polizzi KM, Kontoravdi C (2012) Genome-based kinetic modeling of cytosolic glucose metabolism in industrially relevant cell lines: *Saccharomyces cerevisiae* and Chinese hamster ovary cells. *Bioprocess Biosyst Eng* 35(6):1023–1033. doi:[10.1007/s00449-012-0687-3](https://doi.org/10.1007/s00449-012-0687-3)
 54. Robitaille J, Chen JK, Jolicoeur M (2015) A single dynamic metabolic model can describe mAb producing CHO cell batch and fed-batch cultures on different culture media. *PLoS One* 10(9). doi:[10.1371/journal.pone.0136815](https://doi.org/10.1371/journal.pone.0136815). ARTN e0136815
 55. Villaverde AF, Bongard S, Mauch K, Balsacanto E, Banga JR (2016) Metabolic engineering with multi-objective optimization of

- kinetic models. *J Biotechnol* 222:1–8. doi:10.1016/j.jbiotec.2016.01.005
56. Wang JP, Naik PP, Chen HC, Shi R, Lin CY, Liu J, Shuford CM, Li Q, Sun YH, Tunlaya-Anukit S, Williams CM, Muddiman DC, Ducoste JJ, Sederoff RR, Chiang VL (2014) Complete proteomic-based enzyme reaction and inhibition kinetics reveal how monoclignol biosynthetic enzyme families affect metabolic flux and lignin in *Populus trichocarpa*. *Plant Cell* 26(3):894–914. doi:10.1105/tpc.113.120881
57. Zomorodi AR, Maranas CD (2012) OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput Biol* 8(2):e1002363. doi:10.1371/journal.pcbi.1002363
58. Shrestha UT (2009) Microbial association–microbial interaction. Blogspot. <http://upendrats.blogspot.com/2009/08/microbial-associationmicrobial.html>. Accessed May 31 2016
59. Hansen SK, Rainey PB, Haagenen JAJ, Molin S (2007) Evolution of species interactions in a biofilm community. *Nature* 445(7127):533–536. doi:10.1038/Nature05514
60. Xavier JB (2011) Social interaction in synthetic and natural microbial communities. *Mol Syst Biol* 7. doi:10.1038/Msb.2011.16. Artn 483
61. Fuhrman JA (2009) Microbial community structure and its functional implications. *Nature* 459(7244):193–199. doi:10.1038/Nature08058
62. Wintermute EH, Silver PA (2010) Emergent cooperation in microbial metabolism. *Mol Syst Biol* 6:407. doi:10.1038/msb.2010.66
63. Minty JJ, Singer ME, Scholz SA, Bae CH, Ahn JH, Foster CE, Liao JC, Lin XN (2013) Design and characterization of synthetic fungal-bacterial consortia for direct production of isobutanol from cellulosic biomass. *Proc Natl Acad Sci U S A* 110(36):14592–14597. doi:10.1073/pnas.1218447110
64. Wintermute EH, Silver PA (2010) Dynamics in the mixed microbial concourse. *Genes Dev* 24(23):2603–2614. doi:10.1101/gad.1985210
65. Stolyar S, Van Dien S, Hillesland KL, Pinel N, Lie TJ, Leigh JA, Stahl DA (2007) Metabolic modeling of a mutualistic microbial community. *Mol Syst Biol* 3:92. doi:10.1038/msb4100131
66. Bizukojc M, Dietz D, Sun J, Zeng AP (2010) Metabolic modelling of syntrophic-like growth of a 1,3-propanediol producer, *Clostridium butyricum*, and a methanogenic archaeon, *Methanosarcina mazei*, under anaerobic conditions. *Bioprocess Biosyst Eng* 33(4):507–523. doi:10.1007/s00449-009-0359-0
67. Lewis NE, Schramm G, Bordbar A, Schellenberger J, Andersen MP, Cheng JK, Patel N, Yee A, Lewis RA, Eils R, König R, Palsson BO (2010) Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nat Biotechnol* 28(12):1279–1285. doi:10.1038/nbt.1711
68. Eleftheria Tzamali PP, Tollis IG, Reczko M (2009) Computational identification of bacterial communities. *Int J Biol Biomol Agric Food Biotechnol Eng* 3(4):185–192
69. Nagarajan H, Embree M, Rotaru AE, Shrestha PM, Feist AM, Palsson BO, Lovley DR, Zengler K (2013) Characterization and modelling of interspecies electron transfer mechanisms and microbial community dynamics of a syntrophic association. *Nat Commun* 4:2809. doi:10.1038/ncomms3809
70. Ibarra RU, Fu P, Palsson BO, DiTonno JR, Edwards JS (2003) Quantitative analysis of *Escherichia coli* metabolic phenotypes within the context of phenotypic phase planes. *J Mol Microbiol Biotechnol* 6(2):101–108
71. Tzamali E, Poirazi P, Tollis IG, Reczko M (2011) A computational exploration of bacterial metabolic diversity identifying metabolic interactions and growth-efficient strain communities. *BMC Syst Biol* 5. doi:10.1186/1752-0509-5-167. Artn 167
72. Mo ML, Jamshidi N, Palsson BO (2007) A genome-scale, constraint-based approach to systems biology of human metabolism. *Mol Biosyst* 3(9):598–603. doi:10.1039/b705597h
73. Wahrheit J, Nicolae A, Heinzle E (2011) Eukaryotic metabolism: measuring compartment fluxes. *Biotechnol J* 6(9):1071–1085. doi:10.1002/biot.201100032
74. Borenstein E, Feldman MW (2009) Topological signatures of species interactions in metabolic networks. *J Comput Biol* 16(2):191–200. doi:10.1089/cmb.2008.06TT
75. Freilich S, Kreimer A, Borenstein E, Yosef N, Sharan R, Gophna U, Ruppin E (2009) Metabolic-network-driven analysis of bacterial ecological strategies. *Genome Biol* 10(6):R61. doi:10.1186/gb-2009-10-6-r61
76. Lehmann L, Keller L (2006) The evolution of cooperation and altruism—a general framework and a classification of models. *J Evol Biol* 19(5):1365–1376. doi:10.1111/j.1420-9101.2006.01119.x
77. Nadell CD, Foster KR, Xavier JB (2010) Emergence of spatial structure in cell groups and the evolution of cooperation. *PLoS Comput Biol* 6(3):e1000716. doi:10.1371/journal.pcbi.1000716
78. Shou W, Ram S, Vilar JM (2007) Synthetic cooperation in engineered yeast populations.

- Proc Natl Acad Sci U S A 104(6):1877–1882. doi:[10.1073/pnas.0610575104](https://doi.org/10.1073/pnas.0610575104)
79. Vallino JJ (2003) Modeling microbial consortiums as distributed metabolic networks. *Biol Bull* 204(2):174–179
80. Frey E (2010) Evolutionary game theory: theoretical concepts and applications to microbial communities. *Physica A* 389(20):4265–4298
81. Muller S, Vogt C, Laube M, Harms H, Kleinstaub S (2009) Community dynamics within a bacterial consortium during growth on toluene under sulfate-reducing conditions. *FEMS Microbiol Ecol* 70(3):586–596. doi:[10.1111/j.1574-6941.2009.00768.x](https://doi.org/10.1111/j.1574-6941.2009.00768.x)
82. Zhuang K, Ma E, Lovley DR, Mahadevan R (2012) The design of long-term effective uranium bioremediation strategy using a community metabolic model. *Biotechnol Bioeng* 109(10):2475–2483. doi:[10.1002/bit.24528](https://doi.org/10.1002/bit.24528)
83. Zhuang K, Izallalen M, Mouser P, Richter H, Risso C, Mahadevan R, Lovley DR (2011) Genome-scale dynamic modeling of the competition between *Rhodospirillum rubrum* and *Geobacter* in anoxic subsurface environments. *ISME J* 5(2):305–316. doi:[10.1038/ismej.2010.117](https://doi.org/10.1038/ismej.2010.117)
84. Salimi F, Zhuang K, Mahadevan R (2010) Genome-scale metabolic modeling of a clostridial co-culture for consolidated bioprocessing. *Biotechnol J* 5(7):726–738. doi:[10.1002/biot.201000159](https://doi.org/10.1002/biot.201000159)
85. Hanly TJ, Henson MA (2011) Dynamic flux balance modeling of microbial co-cultures for efficient batch fermentation of glucose and xylose mixtures. *Biotechnol Bioeng* 108(2):376–385. doi:[10.1002/bit.22954](https://doi.org/10.1002/bit.22954)
86. Hanly TJ, Henson MA (2013) Dynamic metabolic modeling of a microaerobic yeast co-culture: predicting and optimizing ethanol production from glucose/xylose mixtures. *Biotechnol Biofuels* 6(1):44. doi:[10.1186/1754-6834-6-44](https://doi.org/10.1186/1754-6834-6-44)
87. Feng X, Xu Y, Chen Y, Tang YJ (2012) Integrating flux balance analysis into kinetic models to decipher the dynamic metabolism of *Shewanella oneidensis* MR-1. *PLoS Comput Biol* 8(2):e1002376. doi:[10.1371/journal.pcbi.1002376](https://doi.org/10.1371/journal.pcbi.1002376)
88. Hoffner K, Harwood SM, Barton PI (2013) A reliable simulator for dynamic flux balance analysis. *Biotechnol Bioeng* 110(3):792–802. doi:[10.1002/bit.24748](https://doi.org/10.1002/bit.24748)
89. Mahadevan R, Edwards JS, Doyle FJ 3rd (2002) Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophys J* 83(3):1331–1340. doi:[10.1016/S0006-3495\(02\)73903-9](https://doi.org/10.1016/S0006-3495(02)73903-9)
90. Zomorodi AR, Islam MM, Maranas CD (2014) d-OptCom: dynamic multi-level and multi-objective metabolic modeling of microbial communities. *ACS Synth Biol* 3(4):247–257. doi:[10.1021/sb4001307](https://doi.org/10.1021/sb4001307)

Extended Metabolic Space Modeling

Pablo Carbonell, Baudoin Delépine, and Jean-Loup Faulon

Abstract

Determining the fraction of the chemical space that can be processed in vivo by using natural and synthetic biology devices is crucial for the development of advanced synthetic biology applications. The extended metabolic space is a coding system based on molecular signatures that enables the derivation of reaction rules for metabolic reactions and the enumeration of all possible substrates and products corresponding to the rules. The extended metabolic space expands capabilities for controlling the production, processing, sensing, and the release of specific molecules in chassis organisms.

Key words Metabolic modeling, Enzyme reactions, Pathways, Products, Chassis

1 Introduction

The set of chemical compounds that organisms can process and synthesize is finite. Such a finite set, however, is not fully known yet. Based on a model that accounts for versatility of enzymatic reactions, we describe here a computational protocol to estimate the extent of such a full metabolic space. The extended metabolic space can be screened to list any possible biological circuit that can be conceived, such as the ones that are used to produce, detect, and process chemicals.

To fully exploit the metabolic space, an essential requirement is having a thorough knowledge of the metabolome associated with any given organism. However, experimental evidences from metabolomics analyses often show that with currently known metabolites one cannot cover the ranges of masses found in actual samples, and consequently there is an impelling need of completing the metabolomes and reactomes of interest for metabolic design [1, 2]. Furthermore, the metabolic phenotype of an organism may vary upon different conditions such as during different growth states leading to variations in the metabolite profile [3]. Besides such sources of uncertainty in samples, many unassigned peaks should be due to promiscuous activities of enzymes not yet

characterized because of the lack of an appropriate description of the mechanisms of enzyme promiscuity.

Our group has addressed the issue of complexity by proposing a tradeoff solution based on molecular signatures [4]. Our molecular signature codes for changes in atom bonding environments where the reaction is taking place. The advantage of the signature method is that the reaction rules describe the changes in the environments of the atoms belonging to the catalytic center of the reactions, and the size of the environment (named diameter) can be tuned to control the combinatorial explosion of possible compounds. Moreover, reaction signatures are robust to unbalanced reactions and can be created automatically without the need of any atom-atom mapping. The signature representation has shown itself to be specially well suited for modeling the mechanisms of enzyme promiscuity [5], paving by these means the way toward engineering innovation in metabolic networks. Either through directed evolution [6] or random selection [7], latent capabilities present in enzymes as modeled by the extended metabolic space can be potentially enhanced to optimize the desired activity and eventually implemented as a biological part containing a metabolic circuit.

Here, we describe the necessary steps to generate an extended metabolic space and how to compute all viable routes within the extended space that determine a viable pathway connecting a desired target to the chassis organism (Fig. 1).

2 Materials

Materials for the described computational protocols consist basically of datasets obtained from public databases and processing software.

- A metabolic database of reference covering chemical structures and reactions. Metanetx [8] is a consensus database that reconciliates multiple databases.
- Models of metabolism for chassis organisms. Biomodels [9] and BiGG [2], among others, are databases containing genome-scale models for most commonly used organisms.
- Software to compute molecular signatures, which are a specialized type of topological chemical descriptors. MolSig [4], among others, is an open-source package that provides such capabilities.
- Matrix manipulation software such as octave, MATLAB, scipy, R, etc.
- Computation of elementary modes. Efmtree [10] provides both a Java and MATLAB-based efficient implementations.
- Software for chemical manipulation. Some of the most popular implementations are RDKit, Marvin, CDK, KNIME (Table 1).

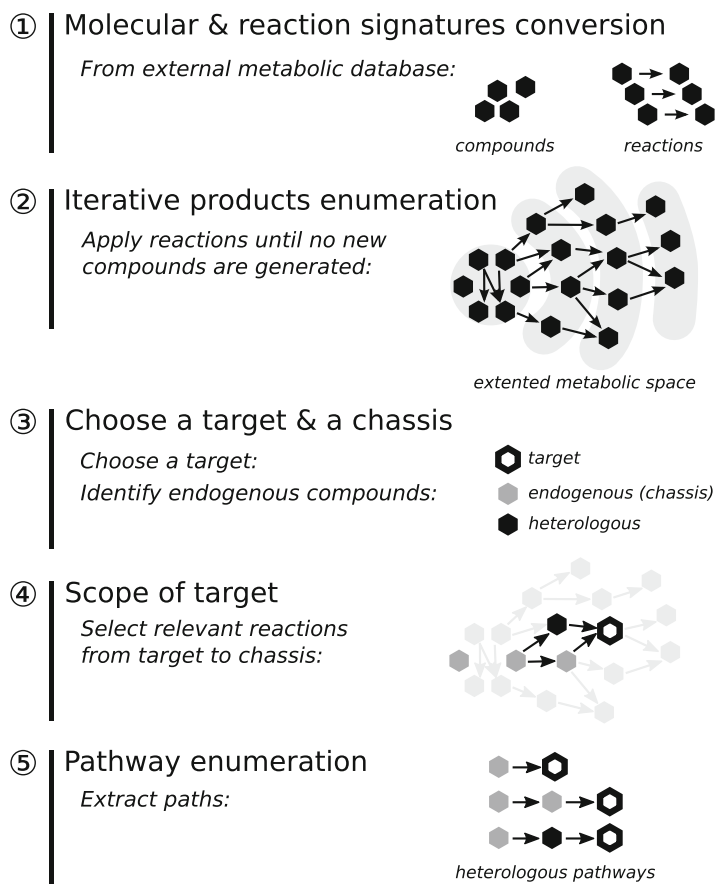


Fig. 1 Steps involved in the construction of the extended metabolic space. The first step consists of converting compounds and reactions into molecular signatures. The second step enumerates new products by an iterative algorithm applied to the reaction signatures. The third step consists of choosing a target, i.e., a reaction or a compound, and a chassis organism. The fourth step determines the metabolic scope linking the chassis to the target. Finally, the fifth step enumerates all viable pathways connecting the chassis to the target

3 Methods

3.1 Computation of Molecular Signatures

The first step to generate an extended metabolic space is to encode all compounds of a metabolic database in a format that will allow the subsequent encoding of enzymatic reactions. We propose here to showcase the important steps that should be kept in mind through the use of one of the available encoding methods, the molecular signature [4] (*see Note 1*).

1. Initially gather compounds from a metabolic database. This database must have structural data for compounds and reactions, and ideally be linked to a whole-cell model (*see Note 2*).

Table 1
A selection of software tools for modeling in the extended metabolic space

| Name | Keyword | Comment |
|-----------------------------|-----------------------------|--|
| <i>Stand-alone software</i> | | |
| Cytoscape | Graph visualization | Cytoscape can be used to manually explore and visualize the EMS [11] |
| efmtool KNIME | Elementary flux Workflow | Computation of elementary flux modes [10] Knime propose to create automatic processes (“workflow”) through a drag-n-drop interface of small tasks (“node”). It is useful for reproducibility of data analysis [12] |
| MarvinSketch | Chemical editor | ChemAxon’s chemical editor. Useful to visualize compounds and reactions, especially to manually inspect difficult cases. URL: http://www.chemaxon.com |
| MolSig | Molecular signatures | Compute molecular signatures from MDL MolFile. URL: http://molsig.sourceforge.net/ |
| <i>Python libraries</i> | | |
| COBRApy | Constraints-based models | A constraint-based steady-state simulation analysis for genome-scale models [13] |
| NetworkX | Graph exploration | NetworkX has an intuitive interface and an extensive documentation. It is a good solution to handle the conversion of the EMS into standard graphs format, or to programmatically explore the EMS. URL: https://networkx.github.io |
| RDKit | Chemoinformatic toolbox | RDKit make it very easy to handle chemical structures, especially to standardize compounds. URL: http://www.rdkit.org |

2. Check compounds for incomplete structural data. Some compounds can be defined with incomplete Markush structure or wildcard atoms. Those compounds typically stand to define classes of compounds (e.g., “an alcohol”) and should be removed since they cannot be interpreted through the molecular signatures algorithm used in this protocol.
3. Standardize compounds. Molecular signatures encode directly molecular graphs from a MDL MolFile input. Users must ensure that compounds (resp., chemical groups) that should be considered identical have the same molecular graph (resp., subgraph) (*see Note 3*).
 - (a) Neutralize or remove charges. As much as possible, chemical groups should be represented with the same protonation state to prevent different tautomeric forms. One can either use heuristics to add or remove hydrogen when necessary or simply remove all charges from the compound dataset.

- (b) Choose one conjugated form by compound. This is particularly important for aromatic compounds, which could appear under different kekulé forms in the database. A good solution is to explicitly use aromatic bonds in the molecular graph description.
 - (c) Use a consistent hydrogen representation, either implicit or explicit.
4. To compute the signature of a chemical compound, we need initially to consider its molecular graph. Let $G(V, E)$ be the molecular graph associated with some chemical compound C and let $a \in V$ ($b \in E$) be an atom (bond) of G . The atomic signature of atom a of diameter d , ${}^d\sigma(a)$, is a canonical representation of the subgraph of G spanned by its vertices at a maximum distance of $d/2$ from a . From a chemical point of view, this corresponds to a circular fragment of the compound centered on a .
 5. The molecular signature of a molecular graph G of diameter d associated with C , ${}^d\sigma(G)$, is defined as the list of all atomic signatures of diameter d (one by atom). Therefore, a molecular signature is a list of overlapping molecular fragments.
 6. Depending on the diameter d , a molecular fingerprint can show degeneracy, i.e., a same molecular signature can represent more than one molecular graph G , much like a chemical formula can correspond to several compounds.
 7. Based on previous definitions, the computation of the molecular signature involves two steps:
 - (a) Choose a diameter to encode enzymatic promiscuity. To some extent, enzymes have the ability to process additional reactants that are structurally similar to the known ones. In a context where it is important to maximize the number of reactions to get more leads, modeling promiscuity can reveal itself to be a critical feature (*see Note 4*). We recommend starting with a diameter of 12 and going lower (down to 4) if no satisfying solution can be found.
 - (b) Compute molecular signatures. The MolSig software [4] computes molecular signatures starting with compounds in MDL MolFiles format, which can be easily retrieved from metabolic and chemical databases or converted from other equivalent formats (*see Note 5*).

3.2 Computation of Reaction Signature

The step following the encoding of compounds is the encoding of reactions into reaction signatures. Reaction signatures should be understood as an exchange of fragments. Unlike other reaction models, reaction signatures do not need any atom-atom mapping to be computed, nor do they need reaction to be balanced (*see Note 6*).

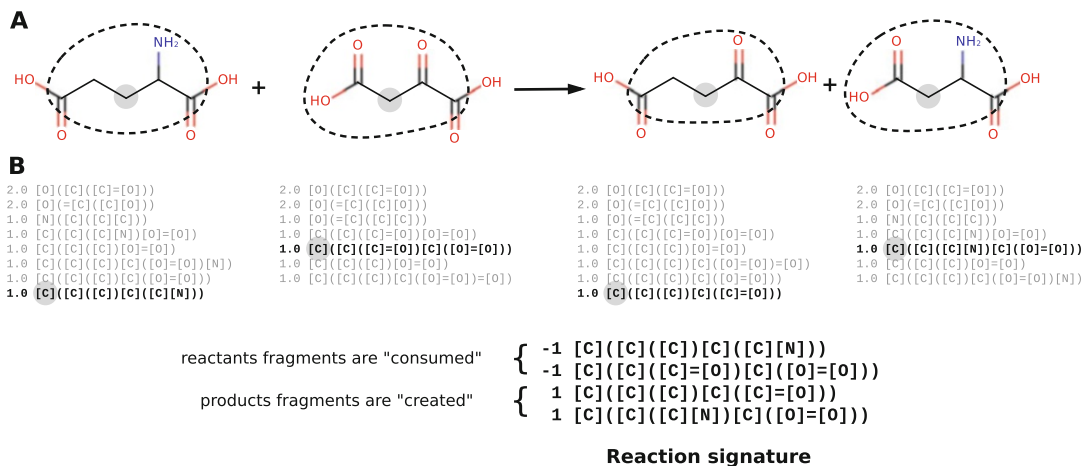


Fig. 2 Reaction signature of an aspartate transaminase (EC 2.6.1.1, $d = 4$). Panel (a) shows the structure of the compounds involved in the original reaction (aspartate + 2-oxoglutarate \rightarrow oxaloacetate + glutamate). Fragments (atomic signature) that are kept in reaction signature are circled (dashed line) and their centers marked by a gray dot. Moieties outside of the circle are allowed to vary under an enzymatic promiscuity hypothesis ($d = 4$). Panel (b) shows the atomic signatures and resulting reaction signature ($d = 4$). Fragments involved in the reaction signature are highlighted (bold) in molecular signatures. Note that several fragments by compound can end up in the reaction signature, even if that is not the case here

1. Let R be a reaction for which all substrates $\{S_i, i \in [1, n]\}$ and products $\{P_j, j \in [1, m]\}$ are encoded in molecular signatures, respectively $\{^d\sigma(S_i), i \in [1, n]\}$ and $\{^d\sigma(P_j), j \in [1, m]\}$. The reaction signature is defined as follows: $^d\sigma(R) = \cup_{j=1}^m ^d\sigma(P_j) - \cup_{i=1}^n ^d\sigma(S_i)$ (see **Note 7**).
2. Thus, $^d\sigma(R)$ is the difference in terms of atomic signatures (i.e., molecular fragments) occurring during a reaction; created (resp. consumed or needed) fragments being positives (resp. negatives). In this context, the diameter d corresponds to the reacting moieties and their neighboring atoms (the environment), hence the possibility to tune the degree of the enzymatic promiscuity hypothesis by increasing or decreasing d (Fig. 2).

3.3 Products Enumeration

Once reactions have been encoded into reaction signatures, they can be applied to compounds to predict potential products under the enzymatic promiscuity hypothesis.

1. Let DB be a database binding compounds signatures to their respective molecular graphs.
2. Let $^d\sigma(R)$ be the molecular signature associated with a reaction R , and $\{S'_i, i \in [1, n]\}$ a set of candidate substrates potentially reacting together.

3. Under the enzymatic promiscuity hypothesis determined by d , we predict that R can process any candidate substrate $\{S_i', i \in [1, n]\}$ if:
 - (a) $\cup_{i=1}^n {}^d\sigma(S_i') \supseteq \{x \in {}^d\sigma(R), x < 0\}$, i.e., if the signatures of candidate substrates include all fragments consumed by R ,
 - (b) and the predicted product(s) signature(s) ${}^d\sigma(P')$ correspond to some previously known compound(s) in DB , with ${}^d\sigma(P') = \cup_{i=1}^n {}^d\sigma(S_i') + {}^d\sigma(R)$ (*see Note 8*).
4. Being able to model enzymatic promiscuity assumes that reaction signatures can be used with other substrates than the ones in the native reaction. In turn, alternative substrates produce new products. Those compounds may be absent from the metabolic space, i.e., the set of known metabolites. Therefore, reaction signatures extend the metabolic space by linking potentially new compounds to the metabolism (*see Note 9*).

3.4 Chassis Modeling in the Extended Metabolic Space

In the previous sections we have described the protocol that allows extending the metabolic space. When the extension is applied to a metabolic network consisting of all known metabolic reactions, we arrive at the full description of all available metabolic capabilities. Some of these capabilities are going to be common to several groups of organisms, such as reactions in the central metabolism, while others like secondary metabolism will be specific to some groups. In applications such a biotechnology, the organism that is engineered is known as the chassis organism and often the objective will be to expand the natural capabilities of the chassis by introducing heterologous enzymes. In this section, we will describe how to model the chassis organism as a subset of the extended metabolic space.

1. The extended metabolic space of diameter d , denoted by M_d , represents all the possible compounds C and allowed transformations (reactions) R between compound as spanned by the enumerated reactions computed by following the method described.
2. A chassis is a subset of the extended metabolic space $O_d \subset M_d$ that corresponds to the extended metabolic network of an organism at signature diameter d . A chassis is defined by the set of nominal reactions annotated for the enzymes present in the organism.
3. The list of nominal metabolic reactions for a given organism can be compiled from databases such as KEGG [14], MetaCyc [15], BiGG [2], BRENDA [16], etc. The choice of one database over the others depends on several factors:
 - (a) The degree of curation of the model.
 - (b) The free and open availability of the model.

- (c) The way the model is going to be analyzed, i.e., network analysis, steady-state simulation or simply as a reference list of metabolites and reactions (*see* **Note 10**).
- 4. In silico organism models showing a good degree of accuracy and reproducibility are currently available for many industrial strains, including *Escherichia coli*, *Saccharomyces cerevisiae*, or *Bacillus subtilis*. They can be generally downloaded in a SBML format [17].
- 5. To determine O_d , each reaction in the reference model is augmented with the set of enumerated reactions of the chassis in the extended metabolic space, resulting in an extended model (*see* **Note 11**).

3.5 Computing the Scope

The next step in modeling in the extended metabolic space is to have an understanding of the design space for a given target metabolic activity. In other words, we want to compute the metabolic scope connecting some target reaction to the chassis. To that end, we provide in this section some relevant definitions and a two-step procedure that allows the determination of the metabolic scope.

1. A minimal pathway is defined as any set of reactions connecting the chassis to the target that are minimal:
 - (a) They form a viable production pathway in terms of precursors availability.
 - (b) All reactions are essential, i.e., the removal of any reaction renders nonviable the pathway (*see* **Note 12**).
2. Based on that definition, the metabolic scope is defined as follows: given an initial set of source metabolites **S** (the chassis) and a final set of target metabolites **T**, the scope is the set of enzymes that are at least involved in one minimal pathway connecting elements of **T** to the source **S**, i.e., the scope should contain only enzymes that are at least essential for establishing one of the metabolic pathways. To compute the scope for a given compound, a two-step procedure can be applied, as described in the following.
3. Reduction of the extended metabolic space to the reachable space of reactions. It consists of the following steps:
 - (a) A compound is defined as reachable if there exists a reachable reaction that can produce it, i.e., a reaction for which all substrates are available.
 - (b) Start from the set of initial compounds **S** and iteratively find newly reachable compounds.
 - (c) The process stops when no new reachable compounds are found.

- (d) Build a graph to keep track of which reactions produced each compound.
4. Backward determination of the scope. It consists of the following steps:
 - (a) Start from the target compound(s) **T**. For each reaction that can produce the target compound(s), add it to the scope.
 - (b) Recursively apply the same procedure on each substrate of the reaction.
 - (c) The recursion stops when initial compounds **S** are reached.

3.6 Enumerating Pathways

Once the extended metabolic scope has been determined, we should be interested in enumerating all viable metabolic pathways connecting the source to the target. This turns out to be a computationally complex problem that can be solved through several approaches [18]. We describe here a solution based on the computation of elementary flux modes [19] (*see Note 13*). EFMs are the set of minimal pathways that are nontrivial solutions to the steady-state equation whose combination can describe any possible path in the network (*see Note 14*).

1. Define the augmented metabolic space formed by the union of the reactions in the chassis and in the scope (Fig. 3a).
2. Construct a stoichiometric matrix **S** where each row corresponds to a compound and each column to a reaction of the previous augmented metabolic spaces and the value of each cell is the stoichiometric coefficient (Fig. 3b).
3. Remove all rows representing initial compounds (*see Note 15*).
4. Remove all rows representing compounds that are produced by a reaction but never used in any other.
5. Merge identical columns by deleting redundant columns and renaming the remaining column with the names of all reactions (*see Note 16*).
6. Add an additional column to create a flux out for the target compound.
7. Several toolboxes exist that allow efficiently computing the elementary modes (Fig. 3c). For instance, *efmtool* [10] provides an efficient implementation that can either run in MATLAB or in Java.
8. Expand resulting elementary modes into the pathway solutions by enumerating all combinations of merged reactions in each elementary mode (Fig. 3d).

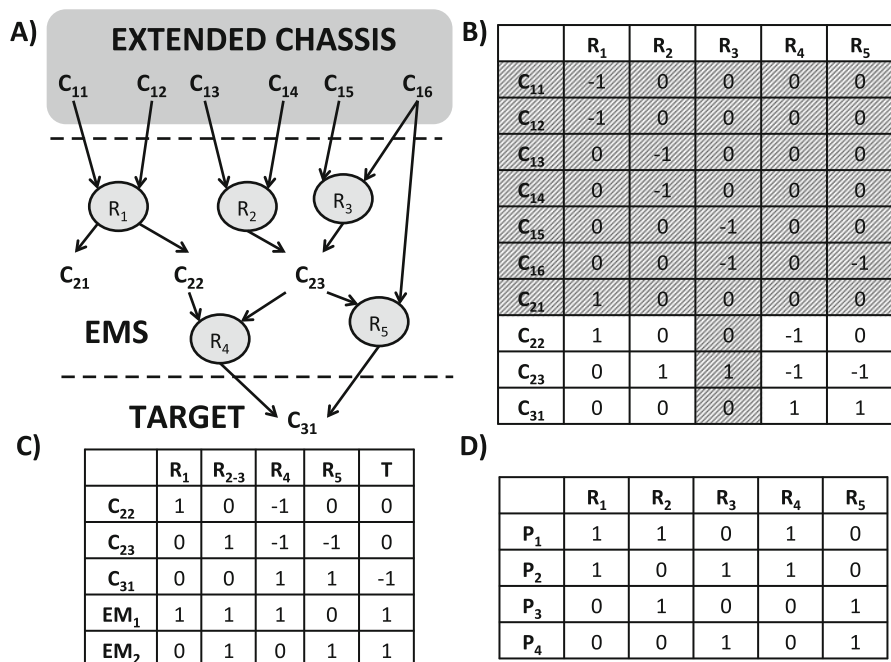


Fig. 3 Example of pathway enumeration in the extended metabolic space. Panel (a) shows the scope graph connecting compounds in the extended chassis (C_{11} , C_{12} , C_{13} , C_{14} , C_{15} , C_{16}) to target compound C_{31} through reactions R_1 , R_2 , R_3 , R_4 and intermediate compounds C_{21} , C_{22} , C_{23} in the extended metabolic space (EMS). Panel (b) displays the equivalent stoichiometric matrix. *Grayed columns and rows* are discarded in the enumeration, as described in the enumeration protocol. Panel (c) shows the reduced matrix used for enumeration, containing an additional reaction T for the selected target compound. The enumeration algorithm found two elementary modes EM_1 and EM_2 . Panel (d) shows the resulting four pathways solution P_1 – P_4 after expansion of topological equivalent reactions. Pathways P_1 and P_2 involve three reactions, while pathways P_3 and P_4 involve two reactions

3.7 Design in the Extended Metabolic Space

We have described in previous sections step-by-step methods that generate extended metabolic spaces for (a) global metabolic capabilities; (b) chassis organisms; (c) organisms augmented with desired target activities. From here, resulting extended models can be used in multiple engineering biology applications, from production of chemicals to their sensing and regulation. Some of the main applications developed to date in extended metabolic spaces include the following:

1. Engineering of heterologous pathways for the production of a desired chemical in a chassis organism. To select enzyme sequences for each enzymatic step in the pathway for the most promising routes in the extended metabolic space, a pathway ranking function needs to be defined. The approach is described in detail in the retrosynthetic RetroPath protocol [20] and a demonstration of the application of such a protocol is shown in the XTMS web service [21].

2. Development of novel biosensors based of metabolic pathways. Metabolic pathways that transform a target compound into a detectable compound allow the expansion of the observable extended metabolic space [22]. Such an application has been demonstrated through the SensiPath web service [23].

4 Notes

1. Molecular signatures are an efficient and intuitive way to model metabolites. They are similar to the well-known Extended Connectivity FingerPrint (ECFP) topological fingerprint, which summarizes compounds in lists of circular molecular fragments.
2. Chemical structures and reactions can be found in multiple formats. Reactions are often defined in a database-specific flat-file where reactants are referenced by their compound identifier. Most of the time, you will find a file in MDL SDF or MOL format binding the compound identifiers to their respective structures. Other interchangeable formats are usually available such as SMILES and InChI. Inter-conversion between formats using standard software such as Open Babel [24] yields to equivalent representations of the compound. A sanity check can help to ensure that they all refer to the same compound. This will eventually filter out wrong annotations.
3. Before being converted into molecular signatures, molecular graphs do not need to represent chemically valid compounds in terms of valence, charges, etc. The important point is that compounds (moieties) that should be considered identical according to the final application share the same molecular graph (subgraph). Of course, those simplifications introduced at the compound encoding step must be kept in mind while interpreting the results.
4. Putative enzymes promiscuity can be modeled through molecular signatures given an appropriate diameter. Obviously, as we lower the diameter, the stronger is the promiscuity hypothesis and the riskier are the predictions.
5. Molecular signatures can take into account stereo-chemistry, which is particularly appealing when working with enzymes. Nonetheless, if stereo information is considered, it is important to ensure that it is available (and valid) for most of the compounds; otherwise, compounds with and without stereo information will be perceived differently through signatures.
6. Metabolic databases contain generally a substantial portion of reactions that are not stoichiometrically balanced. Reaction signatures can be computed for reactions that do not need

strictly balanced input reaction. Nonetheless, working with balanced reactions is always recommended and is a sign of a well-curated database.

7. This mathematical expression simply states that the reaction signature is the set formed by the difference between product signatures and reactions signatures. Intuitively, it can be understood as the chemical groups that are transferred or transformed through the reaction.
8. Multi-substrate reactions are difficult to handle with the proposed equation. Indeed, testing all compounds with a reaction would take N^m tries, where N is the total number of compounds in the database and m the number of substrates anticipated for that reaction. A more practical option is to allow promiscuity for only one substrate at a time, therefore limiting the number of trials to N^*m . A complementary approach is to allow promiscuity only for non-cofactors compounds.
9. This feature is particularly desirable to untap enzymes full potential in metabolic engineering applications since it can find an unexpected synthesis route.
10. There is a basic difference between the information that is required in the model to design heterologous metabolic pathways and to estimate steady-state fluxes. In the former case, the most essential information is the knowledge about the metabolites that are endogenous to the organism and therefore can be used as precursors in the heterologous pathway. In the latter case, the accuracy of the stoichiometric relationship between those reactions that directly influence the pathway is required, while partial knowledge about upstream reactions with low influence into the pathway can be tolerated.
11. The extended metabolic space of the model of an organism provides useful information to discover previously unidentified routes and to fill gaps in present models.
12. Pathway minimality is a heuristic condition based on reducing metabolic burden in the cell (a pathway with a less number of enzymes should be more tolerated by the cell because it potentially imposes less stress).
13. Metabolic networks are formally modeled as hypergraphs for pathway enumeration. Basically, the availability of each substrate is required in the reaction to produce the product. That creates some level of complexity higher than in the classical graph pathway enumeration algorithm. Moreover, standard graph approaches do not consider stoichiometry. The stoichiometric approach, in turn, based on linear algebraic decomposition provides an easier analytic approach.

14. Pathway enumeration based on elementary flux modes can become computationally intractable for highly connected networks such as central metabolism. However, in cases where we want to produce some heterologous compound in a chassis organism, pathways are generally almost linear and the elementary flux mode enumeration remains tractable. The enumeration of elementary flux modes can be also expressed as a dual problem using minimal cut sets.
15. We remove all the initial compounds in the chassis, as we already know that they are available. Products of reactions in the scope consuming the initial compounds will be kept for the enumeration.
16. Identical columns represent routes that are topologically equivalent. In order to make the enumeration algorithm more efficient, we remove duplicated columns. However, for the final enumeration we should list each topologically equivalent reaction as an alternative pathway.

Acknowledgment

The authors acknowledge funding from the BBSRC under grant BB/M017702/1, “Centre for synthetic biology of fine and speciality chemicals.” J.L.F. acknowledges funding provided by the ANR under grant ANR-15-CE21-008. B.D. is affiliated to the Doctoral School Structure et Dynamique des Systèmes Vivant, Université Paris-Saclay.

References

1. Wishart DS, Jewison T, Guo AC et al (2013) HMDB 3.0—the human metabolome database in 2013. *Nucleic Acids Res* 41:D801–D807. doi:[10.1093/nar/gks1065](https://doi.org/10.1093/nar/gks1065)
2. Schellenberger J, Park J, Conrad T, Palsson B (2010) BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* 11:213. doi:[10.1186/1471-2105-11-213](https://doi.org/10.1186/1471-2105-11-213)
3. Tu BP, Mohler RE, Liu JC et al (2007) Cyclic changes in metabolic state during the life of a yeast cell. *Proc Natl Acad Sci U S A* 104: 16886–16891. doi:[10.1073/pnas.0708365104](https://doi.org/10.1073/pnas.0708365104)
4. Carbonell P, Carlsson L, Faulon J-L (2013) Stereo signature molecular descriptor. *J Chem Inf Model* 53:887–897. doi:[10.1021/ci300584r](https://doi.org/10.1021/ci300584r)
5. Carbonell P, Faulon J-L (2010) Molecular signatures-based prediction of enzyme promiscuity. *Bioinformatics* 26:2012–2019. doi:[10.1093/bioinformatics/btq317](https://doi.org/10.1093/bioinformatics/btq317)
6. Dougherty MJ, Arnold FH (2009) Directed evolution: new parts and optimized function. *Curr Opin Biotechnol* 20:486–491. doi:[10.1016/j.copbio.2009.08.005](https://doi.org/10.1016/j.copbio.2009.08.005)
7. Kim J, Kershner JP, Novikov Y et al (2010) Three serendipitous pathways in *E. coli* can bypass a block in pyridoxal-5-phosphate synthesis. *Mol Syst Biol*. doi:[10.1038/msb.2010.88](https://doi.org/10.1038/msb.2010.88)
8. Moretti S, Martin O, Van Du Tran T et al (2016) MetaNetX/MNXref reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res* 44:D523–D526. doi:[10.1093/nar/gkv1117](https://doi.org/10.1093/nar/gkv1117)
9. Chelliah V, Juty N, Ajmera I et al (2014) BioModels: ten-year anniversary. *Nucleic Acids Res* 43:D542–D548. doi:[10.1093/nar/gku1181](https://doi.org/10.1093/nar/gku1181)
10. Terzer M, Stelling J (2008) Large-scale computation of elementary flux modes with bit

- pattern trees. *Bioinformatics* 24:2229–2235. doi:[10.1093/bioinformatics/btn401](https://doi.org/10.1093/bioinformatics/btn401)
11. Shannon P, Markiel A, Ozier O et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504. doi:[10.1101/gr.1239303](https://doi.org/10.1101/gr.1239303)
 12. Berthold MR, Cebron N, Dill F et al (2009) KNIME – the Konstanz information miner. *SIGKDD Explor* 11:26–31. doi:[10.1145/1656274.1656280](https://doi.org/10.1145/1656274.1656280)
 13. Ebrahim A, Lerman JAJ, Palsson BO, Hyduke DR (2013) COBRApy: COntstraints-Based Reconstruction and Analysis for python. *BMC Syst Biol* 7:74. doi:[10.1186/1752-0509-7-74](https://doi.org/10.1186/1752-0509-7-74)
 14. Kanehisa M, Goto S, Sato Y et al (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 40: D109–D114. doi:[10.1093/nar/gkr988](https://doi.org/10.1093/nar/gkr988)
 15. Caspi R, Altman T, Dreher K et al (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 40: D742–D753. doi:[10.1093/nar/gkr1014](https://doi.org/10.1093/nar/gkr1014)
 16. Chang A, Schomburg I, Placzek S et al (2014) BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res.* doi:[10.1093/nar/gku1068](https://doi.org/10.1093/nar/gku1068)
 17. Hucka M, Finney A, Sauro HM et al (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19:524–531
 18. Carbonell P, Fichera D, Pandit S, Faulon JL (2012) Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. *BMC Syst Biol* 6:10. doi:[10.1186/1752-0509-6-10](https://doi.org/10.1186/1752-0509-6-10)
 19. Zanghellini J, Ruckerbauer DE, Hanscho M, Jungreuthmayer C (2013) Elementary flux modes in a nutshell: properties, calculation and applications. *Biotechnol J* 8:1009–1016. doi:[10.1002/biot.201200269](https://doi.org/10.1002/biot.201200269)
 20. Carbonell P, Planson A-GG, Faulon J-LL (2013) Retrosynthetic design of heterologous pathways. *Methods Mol Biol* 985:149–173. doi:[10.1007/978-1-62703-299-5_9](https://doi.org/10.1007/978-1-62703-299-5_9)
 21. Carbonell P, Parutto P, Herisson J et al (2014) XTMS: pathway design in an eXTended metabolic space. *Nucleic Acids Res:W389–W394*. doi:[10.1093/nar/gku362](https://doi.org/10.1093/nar/gku362)
 22. Carbonell P, Parutto P, Baudier C et al (2014) Retropath: automated pipeline for embedded metabolic circuits. *ACS Synth Biol* 3:565–577. doi:[10.1021/sb4001273](https://doi.org/10.1021/sb4001273)
 23. Delpine B, Libis V, Carbonell P, Faulon J-L (2016) SensiPath: computer-aided design of sensing-enabling metabolic pathways. *Nucleic Acids Res.* doi:[10.1093/nar/gkw305](https://doi.org/10.1093/nar/gkw305)
 24. O’Boyle NM, Banck M, James CA et al (2011) Open babel: an open chemical toolbox. *J Cheminform* 3:33. doi:[10.1186/1758-2946-3-33](https://doi.org/10.1186/1758-2946-3-33)

Computational Methods to Assess the Production Potential of Bio-Based Chemicals

Miguel A. Campodonico, Sumesh Sukumara, Adam M. Feist, and Markus J. Herrgård

Abstract

Elevated costs and long implementation times of bio-based processes for producing chemicals represent a bottleneck for moving to a bio-based economy. A prospective analysis able to elucidate economically and technically feasible product targets at early research phases is mandatory. Computational tools can be implemented to explore the biological and technical spectrum of feasibility, while constraining the operational space for desired chemicals. In this chapter, two different computational tools for assessing potential for bio-based production of chemicals from different perspectives are described in detail. The first tool is GEM-Path: an algorithm to compute all structurally possible pathways from one target molecule to the host metabolome. The second tool is a framework for Modeling Sustainable Industrial Chemicals production (MuSIC), which integrates modeling approaches for cellular metabolism, bioreactor design, upstream/downstream processes, and economic impact assessment. Integrating GEM-Path and MuSIC will play a vital role in supporting early phases of research efforts and guide the policy makers with decisions, as we progress toward planning a sustainable chemical industry.

Key words Retrosynthetic pathway design, Techno-economic analysis, Biosustainability

1 Introduction

1.1 Overall Workflow

Based on the production quantity, market price, and formulation, the chemical industry can be segmented into three categories: commodity chemicals, fine chemicals, and specialty chemicals. Currently, approximately 200 major commodity chemicals are being produced and more than 750,000 fine chemicals are being offered globally by several vendors. To produce majority of these chemicals, organic synthesis production routes are pursued in different scales based on petroleum and other fossil-based resources as a feedstock [1]. Fueled by developments in the field of metabolic engineering, an alternative to fossil-based chemicals has emerged. This is the fermentation route, which represents an attractive production option due to the renewable and environmentally benign nature

of the feedstocks utilized, compared to the petrochemical counterpart. However, there exist many challenges to overcome for the fermentation-based processes to be accepted as a long-term solution for sustainable chemical production for most commodity and fine chemicals. A comprehensive prospective analysis able to assess the economical and technical feasibility at early research stages for the evaluation of bio-based process performance is necessary to direct research resources toward most attractive chemical products. Key aspects, which need to be addressed during this feasibility assessment, would be the actual biochemically accessible chemical landscape, process cost, environmental impact, and economic analysis considering long-term risks.

To explore the bioprocess solution space, a general workflow has been proposed and presented in Fig. 1 [2]. In this scheme, a multitude of state-of-the-art modeling approaches, which are able to characterize different scales in the value/technology chain, ranging from the cellular physiology to ecological states were integrated. This framework incorporates various computational methods and tools for quantitatively evaluating the biological, technological, economic, and environmental performance of bio-based chemicals. The workflow is centered around a technology catalogue containing both established petrochemical processes and emerging biochemical processes, constructed using parameters calculated via metabolic, bioreactor, and process simulations. Economic and environmental information such as market size, chemical pricing, and life-cycle inventory data is also incorporated (Fig. 1). Such a framework would represent a cohesive synthesis of the knowledge and perspectives of a spectrum of disciplines, including but not limited to: systems biology, metabolic engineering, process engineering, economics, and environmental engineering.

In this chapter, two key computational tools in the overall workflow presented in Fig. 1 will be fully described. First, to determine the biochemically accessible landscape of possible chemical products, GEM-Path [3] an algorithm able to calculate all structurally possible pathways from one target molecule to the host metabolome is presented. Second, the MuSIC [4, 5] simulation framework that allows multi-scale assessment of the techno-economic and environmental performance of proposed bio-based processes will be described.

1.2 GEM-Path

A key step in developing microbial cell factories is a comprehensive prospective in-silico strain design analysis on the specific strain production capacity for any target chemical [6, 7]. To address this step several computational tools and workflows have been developed, which are able to predict de-novo pathways to target chemicals. Many different approaches have been implemented for this task, and increasing attention has been focused on retrosynthetic algorithms based on Biochemical Reaction Operators

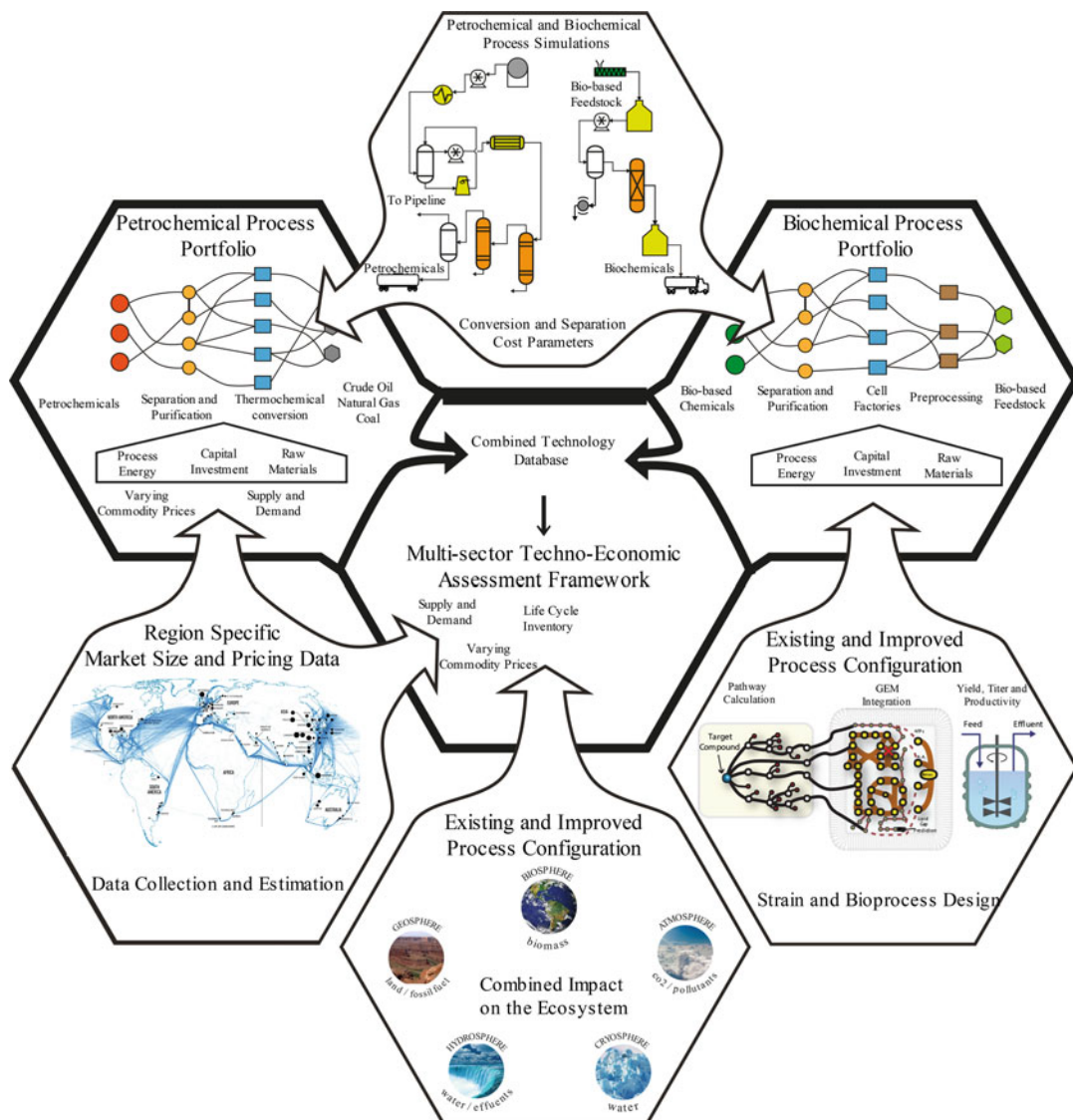


Fig. 1 The concept of a multi-scale, multi-disciplinary assessment framework. The proposed multi-scale, multi-disciplinary assessment framework is centered around a technology catalogue containing both established petrochemical processes and emerging biochemical processes, described as a multi-sector economic network model. This catalogue is constructed using parameters calculated via metabolic, bioreactor, and process simulations. Economic and environmental information such as market size, chemical pricing, and life-cycle inventory data is also incorporated

(BROs) [8–11]. In this chapter, GEM-Path, one of the most comprehensive synthetic pathway prediction algorithms, will be covered and examples of its utility will be illustrated. The GEM-Path pipeline (*see* Fig. 2a) first calculates all structurally possible pathways from one target molecule to the host metabolome. Second, all pathways are integrated into genome scale models (GEMs) and

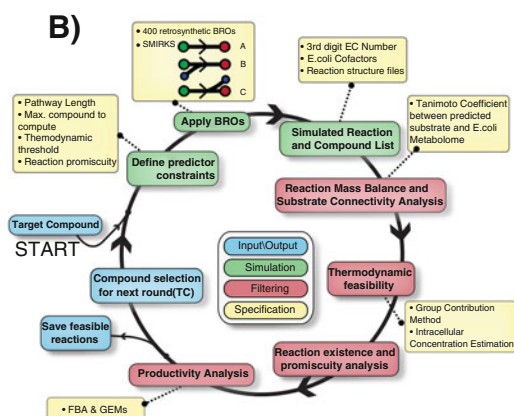
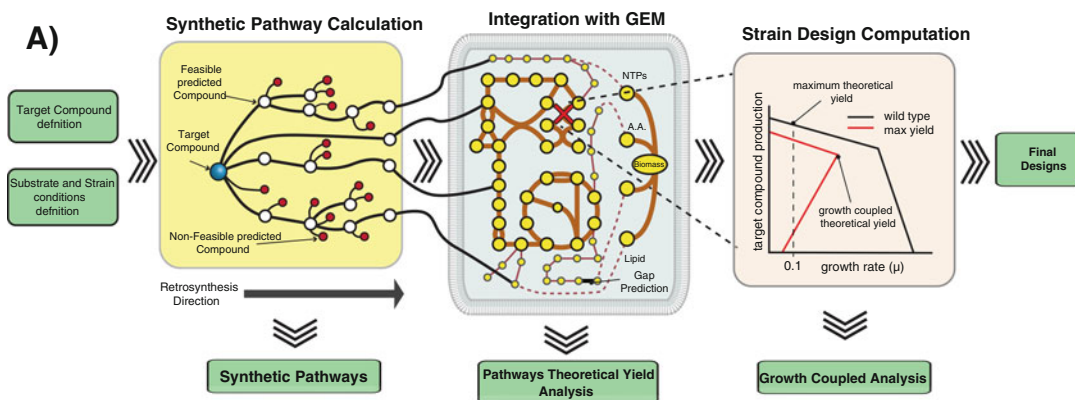


Fig. 2 GEM-Path's general overview. **(a)** Synthetic Pathway Calculation and strain design pipeline: this workflow outlines the integrated process of synthetic pathway prediction (yellow box), constraint-based modeling with the *E. coli* GEM (blue), and strain design computation with design algorithms (pink box). Green boxes represent framework inputs (entry arrows) and general result outputs (exit arrows). From the left, target compounds and substrates/strain conditions were defined to generate synthetic pathways. Synthetic pathways were calculated by using the developed GEM-Path algorithm integrated with GEM computation. Following GEM-Path, each pathway leading to a specific target compound was evaluated for growth-coupled feasibility under previously defined substrate/strain conditions. **(b)** GEM-Path algorithm the iterative loop describes the synthetic pathway calculation process. Colored boxes show process inputs/outputs (blue), synthetic reaction generation (green), and filtering (red) procedures. Furthermore, detailed specifications were assigned to each stage (yellow). Starting with the target compound, BROs were applied and predictor constraints were defined. Mass-balanced generated reactions were filtered in terms of thermodynamic feasibility, reaction existence, and potential promiscuous activity. Then, productivity analysis was performed using FBA. Predicted reactions able to theoretically produce the target compound were saved and non-producing nodes (metabolites) were fed again into the algorithm to continue the pathway generation up to four iterations (reaction steps)

evaluated in terms of their maximum theoretical yield. Finally, strain design computations to evaluate feasibility of growth coupled target production are performed. To generate the set of de-novo pathways, GEM-Path integrates a retrosynthetic algorithm based

on BROs with GEMs into the procedure. The pathway prediction process was developed in an iterative manner (*see* Fig. 2b), which can be broken down into three core steps:

1. Starting from the target metabolite and host organism, predictor constraints are set: such as maximal pathway length, metabolites to compute at each iteration, a thermodynamic threshold, and a reaction promiscuity threshold.
2. Predefined BROs are applied to the target chemical in a retrosynthetic manner for generating the corresponding substrates. All the predicted reactions are then checked for mass balance. Predicted metabolites are structurally compared against host metabolome. If the predicted metabolite matches any compound in the metabolome, flux balance analysis (FBA) [12] is performed to validate the potential production. When performing FBA, a range of substrates can be utilized under aerobic and/or anaerobic conditions.
3. A filtering procedure, which allows thermodynamically feasible reactions, and also reaction existence and promiscuous feasibility, is applied. For maximum theoretical yield estimations, flux balance analysis can be utilized under previously defined media conditions (i.e., carbon sources, nitrogen sources, and oxygenation conditions). Additionally, strain design algorithms able to determine key knock-outs in the metabolism that potentially will allow the growth coupled target chemical production are applied. Since GEM-Path's pathway prediction tool is not freely available, in this chapter the procedures and methods for generating such a tool will be described.

1.3 MuSIC

The heart of the MuSIC framework is a multi-sector network model of a bio-based economy that describes the economic activities within the economy and the exchanges between the economy and the environment. MuSIC currently covers three independent industrial sectors (i.e., agriculture, biochemical, and energy), linked together by a market sector. In each industrial sector, commodities are produced and consumed and subsequently traded inside the market sector that contains the whole industry (*see* Fig. 5). This type of model structure allows easily incorporating other sectors, such as the petrochemical and forestry sectors. The biochemical sector, which specifies the conversion of substrates to chemical products through bio-based processes, is parameterized using a multi-scaled modeling approach that combines cellular metabolism modeling [13, 14], bioreactor design [15–17], and bioprocess modeling [18, 19]. In the biochemical sector, many different combinations of host strains, process technologies, carbon sources, oxygenation conditions, and chemical products can be simulated as different bioprocess entities, thereby enabling a comprehensive performance assessment for a large number of bioprocesses and

chemical products. To mathematically formulate the economic model and formalize the parallelism between metabolism and economy [20], the syntax of a standard constraint-based metabolic model was used. This gives the opportunity to use all previously developed modeling techniques, algorithms, and software tools for constraint-based model and analysis that has been refined during the last decade [13, 14]. Specifically, economic commodities are modeled as metabolites, economic sectors are modeled as compartments, commodity transformation processes are modeled as metabolic reactions, inter-sector trades and environmental emissions are modeled as transport and exchange reactions respectively. Flux balance analysis (FBA) is extensively used to determine the optimal choices of economic pathways, bioprocess technologies, and cell factory designs with respect to a variety of scenarios involving different economic objectives and constraints.

2 Materials

- 2.1 GEM-Path** To get and be able to develop all GEM-Path capabilities, a range of different tools are needed.
- 2.1.1 Programming Platform** Originally, GEM-Path was Developed using MATLAB[®] (The Mathworks Inc.); nevertheless, it can also be extended to other programming languages, such as Python (*see Note 1*).
- 2.1.2 Chemoinformatics Tools** For molecular handling, structural analysis, visualization, and chemical reaction simulations different tools can be applied to the pipeline, such as ChemAxon (ChemAxon Ltd.), OpenBabel [21], CDK [22], and RDKit [23] (*see Note 2*).
- 2.1.3 Biochemical Reaction Operators List** A biochemical reaction operator (BRO) is a generic reaction representation, which is usually represented using the SMARTS [24] notation. Each BRO contains the chemical information to generate every possible compound from a starting substrate. In GEM-Path a total of 443 irreversible BROs were defined based on information collected from the public domain [10, 11, 25]. Approximately, 76% of the reactions in KEGG [26] and 72% of the reactions in BRENDA [27] involved a transformation captured in this defined BRO set.
- 2.1.4 Systems Biology Tools** To evaluate the production efficiency for a specific pathway, a maximum theoretical yield calculation of the target chemical is recommended. GEMs together with FBA represent the most reliable and comprehensive tool to simulate and compare the potential production phenotypes. Highly curated, standardized, and accessible GEMs can be found in the BiGG database models, and

modeling tools such as COBRA Toolbox, COBRAPy, and CAMEO can be utilized to simulate and explore GEMs capabilities. To design a specific strain for the enhanced production of a desired chemical, a range of different tools have been already developed (reviewed in [28]). Furthermore, modeling tools require optimization solvers such as CPLEX, GUROBI, and GLPK to perform calculations.

2.1.5 Databases

Databases are needed for two different purposes: (1) for chemical structure comparison between de-novo predicted chemicals and the host metabolome, and, (2) for synthetic reaction existence analysis and gene assignment. In the first case, the main sources for chemicals in the host metabolome are GEMs. As mentioned above, this can be obtained from BiGG models (*see Note 3*). For the second purpose, databases that hold information regarding metabolites, reactions, genes, and species associations are recommended to utilize such as KEGG, BRENDA, MetaCyc, RHEA, MetaNetX, and MetRxn among others (*see Note 4*). Highly curated databases are recommended to avoid false-positives in the predictions.

2.2 MuSIC

The MuSIC framework and modeling scripts are publicly available as a python package in git [29]. The modeling framework FRAMED (*see Note 5*) is needed to run all associated MuSIC's scripts, which is also publicly available in git [30]. Python and GLPK, as a linear programming solver, are required.

3 Methods

3.1 GEM-Path

GEM-Path's capabilities will be explored using Fig. 2 as a reference. This figure outlines the overall working pipeline and the specific synthetic pathway calculation procedure. Output examples from GEM-Path's original publication [3] are used throughout the methods section.

3.1.1 Target Chemical, Media, and Host Definition

To start GEM-Path procedure, target chemicals, media composition, and microbial host for heterologous productions should be defined. The target chemical should be provided in any chemical file format amenable for the specific chemoinformatics tool in use. Next, to obtain the proper GEM for further calculations, the chassis organisms should be defined (*see Note 6*). Finally, media composition and oxygenation conditions should be set based on physiological data and then integrated into the model (*see Note 7*). Based on the latter, and taking into account the almost negligible time to perform FBA, it is now possible to evaluate and explore a number of different target/substrate/host combinations for any particular project.

3.1.2 Synthetic Pathway Calculation Algorithm

Synthetic Reaction Calculation

The iterative synthetic pathway calculation algorithm can be structured in three main steps (*see* Fig. 2b).

First, to constrain the solution space and obtain the most meaningful pathways in a reasonable amount of time, a maximal pathway length, maximum number of metabolites to iterate after each cycle, and thresholds for analyzing the calculated reaction feasibility need to be set. Next, in a retrosynthetic manner BROs are applied to the target chemical or the set of chemicals selected after each round. To link the predicted reactions with different databases and GEMs, cofactors and EC numbers must be assigned (*see* **Note 8**). The algorithm allows only mass balance reactions to continue the pathway prediction process, and then estimates the structural similarity between each predicted reaction's precursor (*see* **Notes 9** and **10**) and the host metabolome. This step allows the algorithm to further analyze which target chemical to pursue to continue building the pathways toward the closest precursor in the metabolome.

Reaction Feasibility Analysis

After calculating all synthetic reactions, to analyze the potential “in-vivo” reaction feasibility a reaction thermodynamic and existence analysis is performed. For the thermodynamic analysis the ΔrG (kJ/mol) value is an accepted indicator whether the reaction will carry or not flux through the intended direction. Reactions with a ΔrG lower than or equal to 25 kJ/mol are usually defined as feasible reactions (*see* **Note 11**).

Each synthetically calculated reaction needs to be determined whether it exists or not in nature. For this, databases that hold reliable chemical reaction structure information are needed. In GEM-Path, BRENDA was used as a reference and it was processed in way that can be linked with the synthetic reaction predictions (*see* **Note 12**). Furthermore, for synthetic reactions without any match in the BRENDA database, an additional promiscuity analysis was performed (Fig. 3). The first three steps in Fig. 3 describe the main stages for reaction existence and promiscuity analysis. The fourth step in Fig. 3 shows specifically how the promiscuity analysis is performed. First, for a predicted reaction the third level BRENDA EC number database is identified (yellow box). For each reaction in

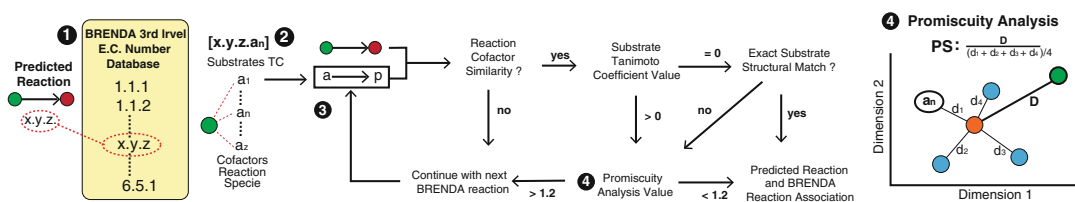


Fig. 3 Reaction existence and promiscuity analysis. The reaction existence comprises four major steps which integrate a cheminformatics-based approach to find the predicted reactions in the BRENDA database. In case no reaction is found in BRENDA, a promiscuity analysis is performed to determine the reaction feasibility

the databases structural information regarding substrates, cofactor uses and species are determined. Second, the predicted reaction substrate (green circle) is compared to the corresponding third level BRENDA EC number database substrates by calculating the TC. From bottom to top, substrate pairs of TCs were sorted in a decreasing order. Third, starting from the lowest TC (a1) until a predicted reaction and BRENDA reaction association is found (an), an iterative decision-making algorithm determines whether the predicted reaction exists in BRENDA or if there is any reaction in the database able to show promiscuous activity. In order to decide whether a reaction might be promiscuous or not, a reaction promiscuity score was calculated based on the similarity between the reaction native substrate and the predicted substrate (Fig. 3, **step 4**). When a specific reaction is sent to promiscuous analysis, nonspecific substrates (blue circles) for the reaction/species association are assigned according to BRENDA databases. By calculating the TC between all of the substrates a reaction promiscuity space was generated. From this space, distances from the centroid for each substrate and promiscuity score were calculated (*see Note 13*).

Target Selection for Next Iteration

Once reaction feasibility is assessed, a pathway connectivity analysis for assessing the heterologous pathway feasibility is performed. Whenever a pathway's precursor matches any metabolite in the host's metabolome, the predicted heterologous pathways are inserted into the host's GEM and FBA is performed to determine the target chemical maximum theoretical yield from different carbon sources (*see Note 14*). A number of different feasibility conditions can be explored during this process, such as substrate utilization and oxygenation conditions. Pathways able to carry flux are stored for further analysis and targets for the next iteration are defined. The targets selection for the next iteration can be done in different ways: (1) selection can be filtered by maximum number of chemicals to process for each iteration. Also, by using different chemical similarity metrics such as the TC (2) chemicals closest to the metabolome, or any specific subset of metabolites can be prioritized. (3) Furthermore, prioritization can be done by using molecular weight and number of carbon atoms as a threshold, or any chemical descriptor, which summarized the properties to target for.

3.1.3 Maximum Theoretical Yield Analysis

Maximum theoretical yield is the most important parameter to pursue experimentally to prioritize pathways. For this purpose, the same procedure used previously (*see* the paragraph "Target Selection for Next Iteration" under Subheading 3.1.2 and **Note 13**) can be utilized to explore the maximum theoretical yield landscape taking into account different variables such as chemical targets, host organism selection, carbon sources, and oxygenation conditions (*see Note 15*).

3.1.4 Strain Design Computations

To implement the advantages of growth coupled target production in any desired organism, several computational methods have been generated to identify the metabolic knockouts necessary to implement coupling of cellular growth to the production of a desired target chemical. Recently, Machado and Herrgård [28] reviewed and compared in detail many available methods proposed so far. The strain design computation workflow can be broken down into two major steps. First, the model has to be preprocessed in a condition-specific manner (*see Note 16*). Model preprocessing is a crucial step in which the size of the model is reduced and the reaction set that could serve as a target for knocking out reactions is defined. This will considerably reduce the computational effort during further strain design calculations. After preprocessing the model, different algorithms can be utilized to perform growth coupled strain design calculations.

3.1.5 Output Example

Taking into account all the previous methods describe above, it is now necessary to visualize and contextualize the results in a comprehensive manner. As an output example Fig. 4 summarizes the results obtained by GEM-Path specifically for the production of 1,3-propanediol by two different computed pathways, pathway #7 that has been experimentally implemented, and pathway #16. Furthermore, the production envelopes for all strain designs of 2–4 gene knockouts are plotted in Fig. 4b. Here, glucose and glycerol were utilized as substrates under aerobic and anaerobic conditions. Solid blue and red lines represent already implemented pathways and “de-novo” GEM-Path calculated pathways, respectively. Moreover, a productivity analysis under different conditions was performed (Fig. 4c), where the shaded areas represent the maximum theoretical production rate by setting the computational minimal growth rate to 0.1 1/h, and the solid areas represent the maximum growth-coupled production rate.

From Fig. 4, many different conclusions can be drawn and are also detailed in the original paper [3]. For example, when comparing the maximum theoretical yield for pathways #16 and #7 under aerobic conditions, an increase of 17% and 25% was observed, and under anaerobic conditions a 6% and 67% increase was observed over glucose, respectively. Also, by using glycerol as a substrate instead of glucose, higher yields were calculated for both aerobic and anaerobic conditions. By using glucose as a substrate the de novo pathways perform slightly better than the already implemented pathway. This means that there might be room for process improvements and also for potential intellectual property generation. Still, this pathway has to overcome the disadvantage of having two reactions predicted to be promiscuous, which is likely to increase time and cost for implementing the pathway.

3.2 MuSIC

The MuSIC framework capabilities will be explored by using the tools available on git, which include the computational procedures utilized for the MuSIC publication [4]. MuSIC is a comprehensive framework that consists of different models at different scales, such as the metabolism, bioreactor, bioprocess, and economy (*see* Fig. 5). The economic model layer is a constraint-based network mathematically formulated as an input-output matrix, which contains the interaction between different economic sectors. In order to model the global economy, it is necessary to parameterize all the sectors comprising the model. Since our aim is to describe and analyze the effect of the biochemical sector in the global economy in detail, first a fully detailed description on how to parameterize this sector will be given. And, next on a more coarse grain type of parameterization, the agriculture and the energy sector will be integrated into the global model. Throughout the methods we will continuously refer to the case study presented in the original publication [4], in which production of 1,3-propanediol (PDO) and 3-hydroxypropionate (3HP) was explored using a combination of different feedstocks, energy sources, host strains, and production pathways. For this analysis, it is necessary to mention that the pathways integrated into the MuSIC framework were previously calculated by using GEM-Path. References to the jupyter computational notebooks that can be found in the main folder when the MuSIC framework is downloaded (*see* **Note 17**), will be given all along the methods according to the next format: First the name of the notebook will be given and separated by hyphen the internal notebook numeration will be added. For example, “S1_metabolism-1” describes the first block or segment in the S1_metabolism notebook.

3.2.1 Biochemical Sector Parameterization

The biochemical sector parameters are calculated via a multi-scale approach involving the utilization of metabolic, bioreactor, and process models. The number of different economic bio-based reactions that will be added to the model will depend on the number of targets, substrates, oxygenation state, host strains, bioreactor, and purification schemes utilized.

Host Metabolism Design and Analysis

To model the metabolism using COBRA methods high-quality GEMs associated with the hosts that are meant to be studied need to be loaded into the MuSic framework (*see* S1_metabolism-1.0 and **Note 18**). Next, the predicted heterologous production pathways to be analyzed have to be defined or the previously generated pathways need to be incorporated into the GEMs (*see* S1_metabolism-1.1 and **Note 19**). At this stage, an analysis on the potential product yield can be performed (*see* S1_metabolism-1.2 and **Note 20**).

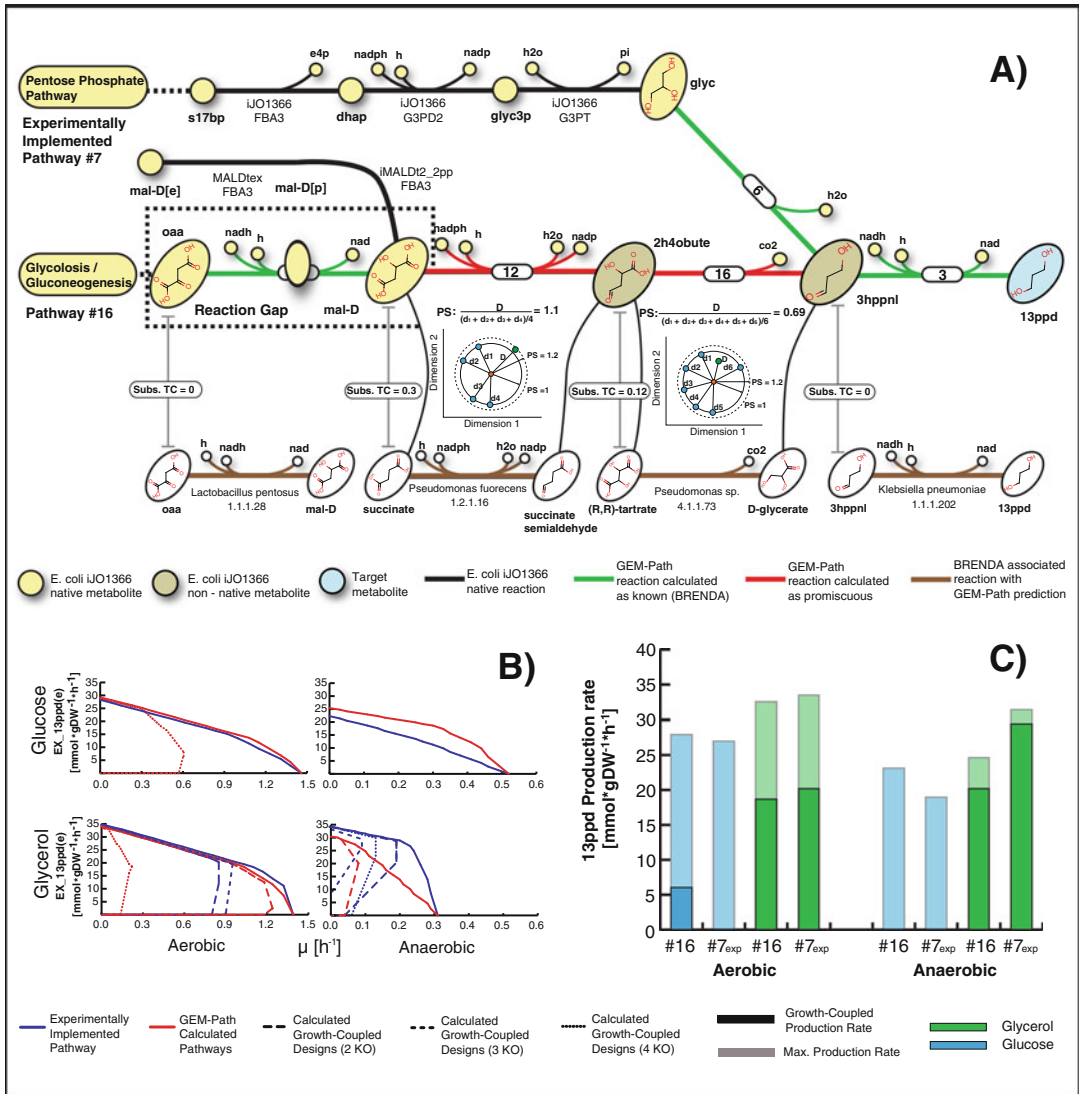


Fig. 4 GEM-Path's results: (a) GEM integrated synthetic pathway calculation (GEM-Path) output for 1,3-propanediol: for each pathway, reactions leading from the host metabolome are shown in *black*. Native and nonnative *E. coli* metabolites are represented in *yellow* and *brown*, respectively. The corresponding target compound is shown in *light blue*. Reactions calculated as known in BRENDA and reaction calculated as promiscuous are shown in *green* and *red*, respectively. For each predicted reaction in pathway #16, specific values of the tanimoto coefficient (TC), promiscuity score (PS), and the corresponding BRENDA reaction (*brown lines*) are shown. For reactions predicted as promiscuous, the corresponding promiscuity space was outlined with the number of metabolites associated with the specific reaction found in BRENDA. For simplicity, a two-dimensional space was plotted, where each of the native BRENDA metabolites (*green*) is separated from the centroid (*red circle*) in 1 dimensionless unit and the predicted metabolite is shown in *green*, with the corresponding distance equal to the PS. A promiscuity score threshold was plotted at a distance equal to 2. Each BRENDA reaction shows the corresponding associated EC number and a species known to catalyze the specific reaction. (b) Strain design production envelopes. (c) Productivity analysis: Results were grouped for

Bioreactor Design and Analysis

Once the GEM and pathways are loaded, the bioreactor type and growth conditions such as substrate utilization and oxygenations states need to be defined and set (*see* S2_bioreactor-2.0). To predict the titer and productivity for each target product/pathway/substrate/oxygenation, a dynamic FBA analysis to simulate batch production is performed (*see* SE_bioreactor-2.1 and **Note 21**). At this stage a bioreactor-scale analysis to investigate the tradeoff between yield, titer, and productivity can be done (*see* S2_bioreactor-2.2).

Bioprocess Design and Analysis

The main goal of these steps is to estimate the process energy cost. In order to do so, it is necessary to first determine the production and purification scheme. Generally, for early stage techno economic and sustainability assessments, a detailed process description is not mandatory (*see* **Note 22**). Also different production and purification schemes for the same substrate/product pair can be computed and analyzed as a part of the overall assessment framework (*see* S3_bioprocess).

3.2.2 Model the Economy

After computing yields, titer, productivities, and energetic costs for all desired product/pathway/feedstock/host/process combinations to parameterize the biochemical industry part of the model, the rest of the sectors are parameterized and the global model is constructed (*see* S4_economy-4.1). Relatively simple models can be used for the non-biochemical sectors, but it is critical that these models include realistic estimates for energy use (e.g., fuel, energy, and land used to produce a ton of glucose in the agriculture sector).

3.2.3 Model Analysis

The comprehensive nature of the model allows exploring the global market solution space by changing a range of different variables and evaluating different objective functions at different model's scales. Here, the case study in the original paper is presented to demonstrate the coverage and type of analysis that can be performed by using the MuSiC framework.

By systematically using FBA, the optimal network flux distribution for producing PDO and 3HP under 24 different scenarios was assessed. Each scenario is a combination of an economic objective and a set of constraints imposed on key variables (e.g., land use), both are specified by a letter and a digit, respectively (*see* Fig. 6).



Fig. 4 (continued) aerobic and anaerobic conditions, associated with the corresponding pathway number. By using glucose (*blue*) and glycerol (*green*) as substrates, maximum theoretical production rate (*shaded bars*) and growth-coupled production rate (*filled bars*) were plotted. FBA was used to determine the maximum theoretical productivity by setting the growth rate to 0.1 1/h and optimizing for the target compound production. Growth-coupled productivity was calculated by knocking out computationally identified reactions and optimizing for growth rate. The maximum value for each condition was reported

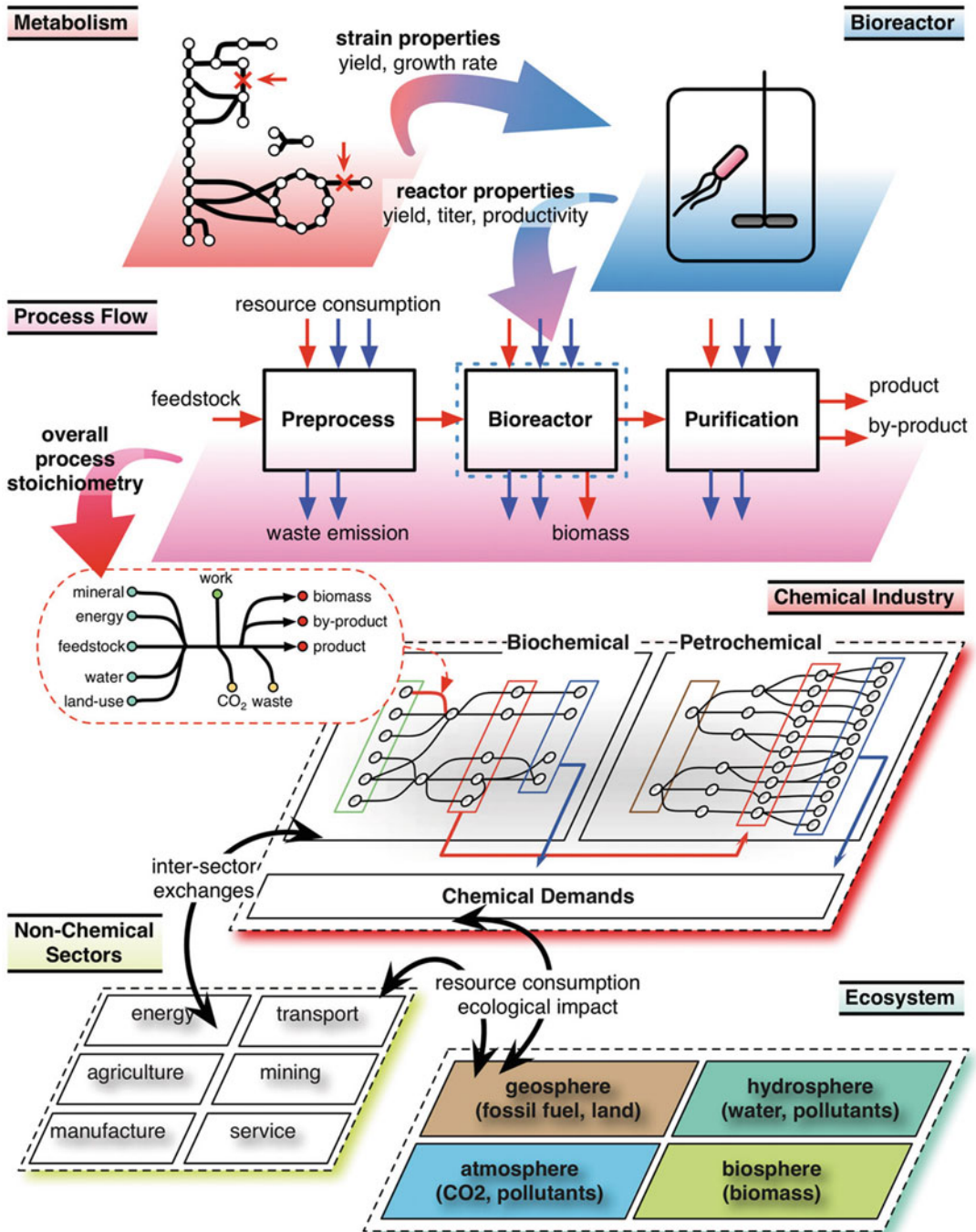


Fig. 5 The proposed MuSIC framework for modeling sustainable biochemical production. This figure illustrates how models at different scales can be integrated into a cohesive multi-scale framework. The metabolic model is used to parameterize the bioreactor model; the bioreactor model is used to parameterize the bioreactor block in the process flow model; each process flow model is abstracted into an overall process stoichiometry and forms a single link in the chemical industry model. The chemical industry model is connected with both other economic sectors and the ecosystem

Analysis on Different Economic Scenarios

For each scenario and each target product, the cash flow and the energy utilization of each sector were determined. Furthermore, the environmental impact using different impact measures and the biomass production were estimated (*see* Fig. 6a). Several conclusions can be drawn from Fig. 6a that can elucidate and aid in prioritizing the best product, feedstock, pathway, host, and feedstock candidates to commercially pursue. In this particular case:

- By looking at all possible scenarios, 3HP production generates more cash flow for the biochemical industry, lower carbon and phosphate emissions, and less cash flow for agricultural industry (Fig. 6a: Row 1) than PDO production per ton of product.
- The cash flow of the biochemical sector does not change significantly when switching from the cash flow maximization objective to the minimization of global warming potential or eutrophication potential (Fig. 6a: Row 1).

Metabolic Engineering Strategies Based on a Comprehensive Multi-Level Analysis

To increase our understanding of the overall system, the MuSIC framework can be utilized to link and compare the different scales or sectors that compose the model. As an example the linking between the bioprocess reactions and the activities of the candidate biosynthetic pathways to a given product can be analyzed (*see* Fig. 6b) to identify pathways that are optimal in different economic scenarios. Furthermore, different strains can be modeled and potential production under different scenarios can be assessed and compared to aid in vivo implementation. For generating Fig. 6b, only 12 distinct pathways were used, suggesting that only a subset of the possible biochemical pathways to a given product is economically feasible. From Fig. 6b many conclusions can be drawn as well, such as:

- Across all scenarios, there is only one pathway that was used in *E. coli* for producing PDO. On the other hand, for producing the same chemical in *S. cerevisiae* there are three viable pathways across all scenarios.
- Although bioreactor simulations suggest that *E. coli* and *S. cerevisiae* performed better under aerobic glucose fed conditions, when analyzing the economy scale simulation, the aerobic glucose-fed condition is never used for *E. coli*, and only used for *S. cerevisiae* in five scenarios that sought to minimize global PO₄e production during PDO production (Fig. 6b). On the other hand, the anaerobic glucose-fed condition is employed ubiquitously for *S. cerevisiae* fermentations (Fig. 6b).
- Anaerobic processes using *S. cerevisiae* as a host would be preferred in the majority of scenarios.

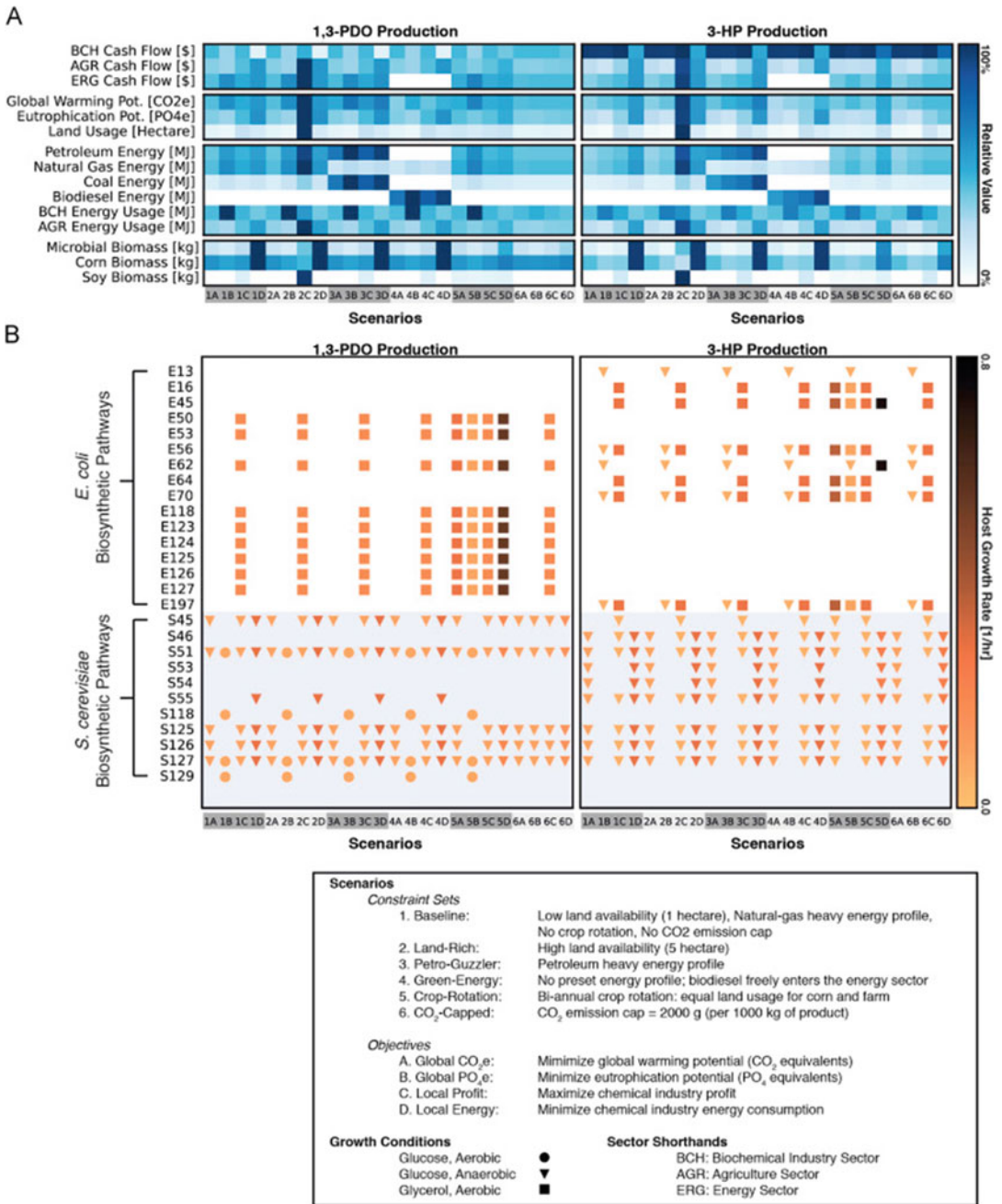


Fig. 6 Economic scenarios, network flows, and metabolic engineering: By using the model, optimal network flow for producing PDO and 3HP for 24 different scenarios representing different political and economic landscapes were evaluated. Each scenario is composed of a design objective and a constraint set, designated with a letter and a number, respectively (see the figure Legend). The top panel (a) shows the predicted sectorial profits, environmental impacts, energy consumptions (by resource type and by sector), as well as the amounts of corn, soy, and microbial biomass produced for each scenario. The bottom panel (b) shows the optimal metabolic engineering decisions—host selection, strain design (growth rate), fermentation condition, and biosynthetic pathways for each scenario

4 Notes

1. To avoid excessive fee charges and get the benefits from a continuous up-to-date software development, open-source programming languages and packages are recommended. Python might be a good starting point, mainly for its user-friendly packages and also because all GEM-Path's capabilities are already available or can be implemented using prebuilt packages.
2. GEM-Path was completely developed using ChemAxon, one of the most comprehensive chemoinformatics tools available.
3. In GEMs, the host metabolome chemical information usually is storage as a chemical name, KEGG ID, PubChem ID, and CHEBI ID among others types of IDs. In case no chemical structure information is provided, we recommend linking these databases programmatically to extract the chemical's information in SMILES, InCHI strings, or any other format that can be utilized by the chemoinformatics tools.
4. It is absolutely necessary to tag these databases to the BRO list to make the synthetic reaction existence evaluation process easier. In GEM-Path EC numbers were utilized for this purpose.
5. When downloading FRAMED for using the MuSIC framework, the kai-brach is mandatory for an adequate algorithm performance.
6. In case no GEM exists for the selected host organism, it is possible to use the closest species available. If there is no available specie, a draft reconstruction might be possible using tools such as ModelSEED [31].
7. Based on experimental data, GEMs usually use as an input the consumption rate of the limiting substrate present in the media, which is most of the time the carbon source or the oxygen uptake. Nitrogen, phosphorous, and inorganic are allowed to freely interact with the media.
8. Each BRO should contain all the information regarding the biochemical reaction, which represents in terms of EC numbers and cofactors. This assignation is crucial for further reaction feasibility analyzing and GEMS integration.
9. For comparing chemical entities first their molecular fingerprints should be calculated and then compared by using different metrics, such as the Tanimoto Coefficient (TC). A molecular fingerprint is a simple record of the fragments (chemical substructures) presented in a chemical structure. It is usually represented as a binary bit string, which can be compared against others by using the TC. The TC basically

determines how similar two fingerprints are. Values are non-negative numbers, where a zero value indicates that the two fingerprints are identical, and the larger the TC value the greater the difference between the two fingerprints.

10. Since the molecular fingerprint only represents the presence or absence of a given bond pattern, when calculating a TC equal to zero for a couple of molecules it actually means that both the molecules share the same bonds according to the fragmentation procedure. This not necessarily means that both the molecules are the same. In those cases, an exact topology match analysis is needed.
11. The ΔrG threshold of 25 kJ/mol is defined based on the estimated variability calculated elsewhere [32], which identify the range of action of metabolic reactions in *E. coli*.
12. The BRENDA database was restructured by lumping together all reactions with the same third level EC numbers. Each level contains known biochemical reactions with the corresponding four-digit EC number association, reaction-organism association, and substrate structure file. The third-level EC number association for the predicted reaction facilitates the identification of the third-level EC class BRENDA subgroup for substrate comparison.
13. The reaction promiscuity score was calculated and analyzed by using *E. coli*'s promiscuous reaction information from iJO1366. Based on the previous analysis, the reaction promiscuity score threshold was set to 1.2 [3].
14. The maximum theoretical yield calculation can be executed by: (1) integrating the heterologous pathways into the GEM, (2) defining the substrate of interest and setting its uptake rate to physiological conditions, (3) setting the oxygen uptake rate to zero for anaerobic conditions or at any physiological value greater than zero for aerobic conditions, (4) setting a minimal growth rate to sustain growth (usually 0.1 1/h), and (5) using FBA to maximize the flux through the exchange reaction associated with the target compound.
15. Usually, in this type of analysis, many different variables are utilized for calculating the maximum theoretical yield. In order to compare the outcome of the simulation accurately, results should be reported in terms of carbon-mol/carbon_mol. Furthermore, bounds by products and carbon dioxide should be accurately constrained to avoid incorrect yield estimations.
16. Model preprocessing has to perform independently for each physiological condition that is meant to be analysis. This means that substrate utilization, oxygenation, and heterologous pathway insertions must be properly constrained before preprocessing the model.

17. Jupyter notebooks can be found when downloading the MuSIC framework from git [29].
18. COBRA models can be downloaded from the BiGG [33] database. This database contains highly curated and reliable GEMs.
19. Pathways can be already known or “de-novo.” The reactions present in the pathways need to be curated and metabolites named based on the GEM nomenclature that is being used.
20. Usually to outline the maximum theoretical performance for a particular target, an envelope analysis is performed to analyze the tradeoff between the strain’s production capacity and growth rate.
21. In order to obtain the titer and productivity for a potential strain, dynamic FBA (dFBA) is utilized extensively [5].
22. For the bioprocess model a simplified flow sheet containing three generic blocks such as bioreactor fermentation, separation of biomass from liquid via centrifugation, and purification of product from by-products via evaporation will be enough for an early-stage phase techno-economic assessment.

References

1. Pollak P (2011) *Fine chemicals: the industry and the business*. John Wiley & Sons, Hoboken, NJ
2. Herrgard M, Sukumara S, Campodonico M et al (2015) A multi-scale, multi-disciplinary approach for assessing the technological, economic and environmental performance of bio-based chemicals. *Biochem Soc Trans* 43:1151–1156
3. Campodonico MA, Andrews BA, Asenjo JA et al (2014) Generation of an atlas for commodity chemical production in *Escherichia coli* and a novel pathway prediction algorithm, GEM-Path. *Metab Eng* 25:140–158
4. Zhuang KH, Herrgård MJ (2015) Multi-scale exploration of the technical, economic, and environmental dimensions of bio-based chemical production. *Metab Eng* 31:1–12
5. Zhuang K, Bakshi BR, Herrgård MJ (2013) Multi-scale modeling for sustainable chemical production. *Biotechnol J* 8:973–984
6. Dai Z, Nielsen J (2015) Advancing metabolic engineering through systems biology of industrial microorganisms. *Curr Opin Biotechnol* 36:8–15
7. Hadadi N, Hatzimanikatis V (2015) Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. *Curr Opin Chem Biol* 28:99–104
8. Carbonell P, Planson A-G, Fichera D et al (2011) A retrosynthetic biology approach to metabolic pathway design for therapeutic production. *BMC Syst Biol* 5:122
9. Cho A, Yun H, Park JH et al (2010) Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Syst Biol* 4:35
10. Henry CS, Broadbelt LJ, Hatzimanikatis V (2010) Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. *Biotechnol Bioeng* 106:462–473
11. Yim H, Harry Y, Robert H et al (2011) Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nat Chem Biol* 7:445–452
12. Orth JD, Ines T, Palsson BØ (2010) What is flux balance analysis? *Nat Biotechnol* 28:245–248
13. Bordbar A, Monk JM, King ZA et al (2014) Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet* 15:107–120
14. McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. *Mol Syst Biol* 9:661
15. Mahadevan R, Edwards JS, Doyle FJ 3rd (2002) Dynamic flux balance analysis of diauxic

- growth in *Escherichia coli*. *Biophys J* 83:1331–1340
16. Zhuang K, Kai Z, Laurence Y et al (2013) Dynamic strain scanning optimization: an efficient strain design strategy for balanced yield, titer, and productivity. *DySScO strategy for strain design. BMC Biotechnol* 13:8
 17. Varma A, Palsson BO (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
 18. Caspeta L, Luis C, Jens N (2013) Economic and environmental impacts of microbial biodiesel. *Nat Biotechnol* 31:789–793
 19. Hermann BG, Patel M (2007) Today's and tomorrow's bio-based bulk chemicals from white biotechnology. *Appl Biochem Biotechnol* 136:361–388
 20. Smart B (1992) Industry as a metabolic activity. *Proc Natl Acad Sci U S A* 89:804–806
 21. O'Boyle NM, Banck M, James CA et al (2011) Open babel: an open chemical toolbox. *J Chem* 3:33
 22. Steinbeck C, Han Y, Kuhn S et al (2003) The chemistry development kit (CDK): an open-source java library for chemo- and bioinformatics. *J Chem Inf Comput Sci* 43:493–500
 23. G. Landrum RDKit. <http://www.rdkit.org>.
 24. James CA, Weininger D, Delany J (1995) Daylight theory manual. Daylight Chemical Information Systems, Irvine, CA
 25. Mu F, Unkefer CJ, Unkefer PJ et al (2011) Prediction of metabolic reactions based on atomic and molecular properties of small-molecule compounds. *Bioinformatics* 27:1537–1545
 26. Kanehisa M (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 34:D354–D357
 27. Curran KA, Alper HS (2012) Expanding the chemical palate of cells by combining systems biology and metabolic engineering. *Metab Eng* 14:289–297
 28. Machado D, Daniel M, Herrgård MJ (2015) Co-evolution of strain design methods based on flux balance and elementary mode analysis. *Metab Eng Commun* 2:85–92
 29. biosustain biosustain/MuSIC-PDO-3HP. <https://github.com/biosustain/MuSIC-PDO-3HP>
 30. cdanielmachado cdanielmachado/framed. <https://github.com/cdanielmachado/framed>
 31. Overbeek R (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33:5691–5702
 32. Henry CS, Broadbelt LJ, Hatzimanikatis V (2007) Thermodynamics-based metabolic flux analysis. *Biophys J* 92:1792–1805
 33. King ZA, Lu J, Dräger A et al (2016) BiGG models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res* 44:D515–D522

Part II

Building and Refactoring Synthetic Metabolic Pathways

Multiplex Genome Editing in *Escherichia coli*

Sheila Ingemann Jensen and Alex Toftgaard Nielsen

Abstract

Lambda Red recombineering is an easy and efficient method for generating genetic modifications in *Escherichia coli*. For gene deletions, lambda Red recombineering is combined with the use of selectable markers, which are removed through the action of, e.g., flippase (Flp) recombinase. This PCR-based engineering method has also been applied to a number of other bacteria. In this chapter, we describe a recently developed one plasmid-based method as well as the use of a strain with genomically integrated recombineering genes, which significantly speeds up the engineering of strains with multiple genomic alterations.

Keywords Synthetic biology, Industrial biotechnology, Lambda Red recombineering, Flippase recombinase, pSIJ8, Genome editing

1 Introduction

Synthetic biology and metabolic engineering require the ability to alter the genome of the strain in question. Often multiple genomic alterations are needed to redirect fluxes, increase tolerance to various conditions, and investigate combinatorial physiological interactions. The most widely used method to generate genomic deletions and insertions in *Escherichia coli* is the use of lambda Red recombineering in combination with an appropriate selection marker, e.g., flippase recombinase to remove the integrated selection markers once the desired alteration has been verified. During lambda Red recombineering, the Gam protein interacts with RecBCD and SbcCD nucleases [1–3], which minimizes the degradation of the introduced double-stranded (ds) DNA. The Exo protein degrades the introduced dsDNA leaving behind a single-stranded (ss) DNA [4, 5], whereas the Beta protein binds to the ssDNA generated by the Exo protein and enables its incorporation into genomic DNA during replication [6–8]. A variety of factors affect dsDNA recombineering efficiency such as transformation efficiency, growth phase, induction times, dsDNA concentration,

and length of the homology arms [9–11]. The use of oligoes containing either phosphorothioate bonds or phosphorylation has furthermore been shown to increase dsDNA recombineering efficiency [12–14]. The genomic location of genes is furthermore important when performing multiple alterations during one round of electroporation [14]. The removal of the selection markers is often done using FRT-flanked antibiotic cassettes combined with flippase (FLP) recombinase [15, 16]. FLP recognizes the integrated 34 bp (minimum) target site (FRT) and removes the selection marker through recombineering of the integrated FRT sites, leaving behind a scar containing a single FRT site [17–19]. Multiple scars left behind in the genome can, when FLP recombinase is present, lead to undesired recombineering events. Since prolonged induction times can in this way affect successful engineering of strains with multiple genomic alterations, it is desirable to keep the engineering steps short. Here, we describe in detail a recently developed one plasmid method, which facilitates the generation of strains with multiple genomic alterations without the need for continuous plasmid re-transformation. We further describe in detail the generation and use of strains with genomically integrated recombineering genes, which significantly speeds up the engineering of strains with multiple genomic alterations, and which facilitates multiple gene deletions during one round of electroporation [14].

2 Materials

2.1 Strains and Plasmids

pKD3 (Cm), pKD4 (Km) [16], pSIJ8 (Ap), pSIJ196 (Sp), pSIJ197 (Gm), and strain SIJ488 [14].

2.2 Media and Reagents

1. LB: 10 g/L tryptone, 5.0 g/L yeast extract, 5 g/L NaCl, dissolve in water and autoclave. For LB-agar plates add 1.5% agar.
2. Antibiotics stock solutions: Ampicillin (Ap) 100 mg/mL, Kanamycin (Km) 50 mg/mL, Spectinomycin (Sp) 50 mg/mL, Gentamycin (Gm) 10 mg/mL, and Chloramphenicol (Cm) 30 mg/mL. Filter sterilize and store at -20°C .
3. Antibiotics working concentration: Ap 100 $\mu\text{g}/\text{mL}$, Km 30–50 $\mu\text{g}/\text{mL}$, Sp 50 $\mu\text{g}/\text{mL}$, Gm 10 $\mu\text{g}/\text{mL}$, and Cm 20–30 $\mu\text{g}/\text{mL}$.
4. L-arabinose and L-rhamnose: 1 M in MilliQ water (Mq-water), filter sterilize. Long-term storage at -20°C , short-term storage (~ 2 months) at 4°C . Use at a final concentration of 15 mM (arabinose) and 50 mM (rhamnose).
5. Ice-cold autoclaved Mq-water.

6. For plasmid preparations and PCR purifications one can use a number of commercially available plasmid and PCR cleanup kits. Primers and reagents for PCRs are further commercially available.

3 Methods

3.1 General Methods

3.1.1 Oligo Design

Design oligoes with 50–100 bp flanking homologies up- and down-stream of the genes to be deleted. 40 bp may be sufficient in some cases. For gene deletions in *E. coli* K-12 MG1655, it is possible to use the flanking homologies that were used to make the Keio-collection [20, 21], and the following sequences for the forward and reverse oligoes to target *frt*-flanked antibiotic resistance cassettes:

Ant-fw N50- GTG TAG GCT GGA GCT GCT TC
 Ant-rv N50- CAT ATG AAT ATC CTC CTT AGT TCC

For other *E. coli* strains it is necessary to check the genome sequence of the respective strain using publically available genome databases such as, e.g., the Kegg database (www.genome.jp) [22] for the design of appropriate homology extensions.

For colony PCRs design ~20 bp oligoes 200–250 bp upstream and/or downstream of the gene to be deleted.

Phosphorothioate bonds can be used to protect dsDNA from exonuclease degradation and thereby enhance dsDNA recombineering efficiency [12–14]. The phosphorothioated modifications should be in the oligo that ensures protection of the lagging strand-targeting dsDNA, and the genomic location and orientation in respect to the origin of replication have to be taken into account. Ecogene (<http://www.ecogene.org>) [23] can be used as an excellent resource for the design of these types of oligoes. For examples of such primers please *see* Jensen et al. [14].

3.1.2 PCRs

For colony PCRs and PCRs using short oligoes targeting, e.g., the Keio collection, standard PCR procedures and PCR programs can be used following the manufacturer's instructions.

For PCRs with long extensions it is possible to use the following dual step PCR program to minimize primer dimer formation (applicable when using the above-mentioned ant-fw and ant-rv targeting primers and Phusion Hot Start II DNA polymerase). It is recommended to target gDNA with genomically integrated *frt*-flanked antibiotic cassettes, or at least *DpnI* treat the PCR product prior to further use to minimize false positives during the gene deletion procedures described below.

Dual step PCR program for primers with long extensions

1. 98 °C for 3 min.
2. 97 °C for 15 s.
3. 60 °C for 30 s.
4. 72 °C for 1 min 30 s.
5. Go to **step 2** for nine more times.
6. 97 °C for 10 s.
7. 65 °C for 30 s.
8. 72 °C for 1 min 30 s.
9. Go to **step 6** for 19 more times.
10. 72 °C for 5 min.
11. 15 °C forever.

3.1.3 Preparation of Electrocompetent Cells

1. Grow cells in LB until desired OD (the final amount of water corresponds to 50 mL cells of OD ~0.4–0.6).
2. Place cells on ice for 5–10 min.
3. Spin down cells at 4000 × *g* (6500 × *g* for MG1655), 4 °C for 7 min.
4. Pour out supernatant and add ~5 mL ice-cold Mq-water (or 10% glycerol), and dissolve pellet by shaking (keep cold). Fill up to 50 mL with precooled water (or 10% glycerol).
5. Spin down cells at 4000 × *g* (6500 × *g* for MG1655), 4 °C for 7 min.
6. Pour off supernatant and dissolve pellet **gently** in 1 mL water (or 10% glycerol).
7. Transfer to 1.5 mL Eppendorf tubes.
8. Spin down cells at 4000 × *g* (6500 × *g* for MG1655), 4 °C for 2 min.
9. Pipet off supernatant and dissolve pellet **gently** in 200 µL water (or 10% glycerol).
10. Place on ice.
11. Aliquot 50 µL electro competent cells into precooled 1.5 mL Eppendorf tubes, keep on ice.

3.2 Transformation of pSIJ8 into Strain of Interest (See Note 1)

1. Grow up cells in 5 mL LB to an OD ~1.0 or overnight.
2. Prepare electrocompetent cells as described above.
3. Add 2.5 µL plasmid by placing the pipette tip in the middle of the cell pellet and stir gently a couple of times. Do not pipet up and down, and make sure to keep the cells cold.
4. Transfer electrocompetent cells with added plasmid to precooled electroporation cuvettes.

5. Electroporate at 1.8 kV ~5 ms.
6. Add 1 mL LB **immediately** and transfer back to 1.5 mL Eppendorf tube. NB: do this before electroporating other cells to be transformed.
7. Place the Eppendorf tube at 30 °C, 250 rpm for 1 h.
8. Plate on LB_{Ap} and grow overnight at 30 °C.

3.3 Transfer of Integrated Recombineering Genes from Strain SIJ488

Integration of the recombineering genes into the genome of *E. coli* K-12 MG1655 strains, and other *E. coli* K-12 strains with appropriate homology regions can be done using pSIJ214 as described elsewhere [14]; however, we recommend transferring the system using either P1-vir transduction (combined with the deletion of the integrated xylSpm-IsceI genes as described in Subheading 3.3.1, step 2) or the following protocol for the transfer of the system.

3.3.1 Transfer by Recombineering

1. Delete the integrated xylSpm-IsceI genes AND the flippase recombinase genes in strain SIJ488 by following the protocol described for deletions below and using the following primers for PCR amplification of an *frt*-flanked Cm-cassette:
 - (a) SIJ488-1fw
 tccgttacaagcgaggctgggtatttcccggcctttctgttatccgaaaGTGT
 A GGCTGGAGCTGCTTC
 - (b) SIJ488-1&2rv
 aaaacataacaggaagaaaaatgccccgcttacgcagggcatccatttaCATA
 TGAATATCCTCCTTAGTTCC
2. Delete the integrated xylSpm-IsceI genes in strain SIJ488 by following the protocol described below and using the following primers for PCR amplification of an *frt*-flanked Km-cassette:
 - (a) SIJ488-2fw
 tccagatggagtcttctgaggctcattactggatctatcaacaggagtcgaag GT
 GTAGGCTGGAGCTGCTTC
 - (b) SIJ488-1&2rv
 aaaacataacaggaagaaaaatgccccgcttacgcagggcatccatttaCATA
 TGAATATCCTCCTTAGTTCC
3. Amplify the lambda Red recombineering genes and the integrated chloramphenicol cassette in strain SIJ488Cm (generated in Subheading 3.3.1, step 1) with 50 bp flanking extensions corresponding to the genomic sequence of the strain of interest. The following targeting oligoes can be used for this purpose:
 - (a) SIJint-2fw N50-TCTAGGGCGGCGGATTTG
 - (b) SIJint-1&2rv N50-CATATGAATATCCTCCTTAGTTCC
4. Amplify the flippase recombinase and the integrated kanamycin cassette in strain SIJ488Km (generated in Subheading 3.3.1,

step 2) with 50 bp flanking extensions corresponding to the genomic sequence of the strain of interest. The following targeting oligoes can be used for this purpose:

(a) SIJint-1fw TCCGTTACAAAGCGAGGCTG

(b) SIJint-1&2rv N50-CATATGAATATCCTCCTTAGTTCC

5. Introduce pSIJ8 into strain of interest as described above.
6. Integrate the lambda Red recombineering genes + Cm-cassette using the same protocol as described for deletions below (you can grow the strain at 37 °C during recovery to start losing pSIJ8).
7. Integrate the flippase recombinase + Km-cassette using the protocol described below for deletions. Note that the Cm-cassette integrated in Subheading 3.3.1, **step 6** will be deleted during this step.
8. Induce the now integrated flippase recombinase and remove the Km-cassette as described below.
9. Check that pSIJ8 has been removed from the strain by restreaking on LB (positive) and LB_{AP} plates (negative control).
10. Verify the integration by colony PCR and sequencing.

3.4 Generation of Multiple Deletion Strains Using pSIJ8 (See Also Notes 1–6)

3.4.1 First Gene Deletion

1. Pick a colony of *E. coli* with pSIJ8 and grow in 250 mL shake flasks with shaking in 50 mL LB_{AP} until OD₆₀₀ ~0.2–0.3.
2. Add 750–1000 µL 1 M filter-sterilized arabinose (to induce the lambda Red recombineering genes).
3. Grow for 30–45 min.
4. Pour cells into 50 mL falcon tubes and place on ice for 5–10 min.
5. Prepare electrocompetent cells as described above.
6. Add 5 µL PCR product to the middle of the cell pellet and stir gently a couple of times with the pipet tip, do not pipet up and down).
7. Transfer electrocompetent cells with PCR product to pre-cooled electroporation cuvettes.
8. Electroporate at 1.8 kV ~5 ms.
9. Add 1 mL LB **immediately** after electroporation and transfer back to 1.5 mL Eppendorf tube.
10. Grow cells with shaking at 30 °C for 1.5–2 h.
11. Plate on appropriate selection plate (e.g., km + 75 µL LB_{AP} spread on plate) and grow overnight at 30 °C.

3.4.2 Removal of Antibiotic Cassette

1. Transfer ~2–8 colonies to, e.g., individual PCR tubes with 20 µL LB.
2. Use 1–2 µL to perform colony PCRs using appropriate primers.

3. While colony PCR is running, grow cell cultures in 1 mL LB w. appropriate antibiotic (e.g., Ap + Km) at 30 °C with shaking.
4. After 2–3 h growth, spin down a verified pre-grown culture at $6500 \times g$ for 2 min. Remove the supernatant and dissolve pellet in 1 mL LB_{AP} + 50 mM rhamnose (start OD ~0.1–0.2).
5. Grow with induction for 4–6 h and streak on LB_{AP} plates for overnight growth at 30 °C.

3.4.3 Second Gene Deletion

1. Transfer ~2–8 colonies to, e.g., individual PCR tubes with 20 µL LB.
2. Do colony PCRs using appropriate upstream and downstream primers.
3. While colony PCR is running, transfer colonies to 1 mL LB_{AP} and grow at 30 °C.
4. Transfer culture with verified removal of cassette to 50 mL LB_{AP} and continue as described above for first gene deletion.

3.4.4 Removal of pSIJ8

1. Pick a few colonies and grow the strains individually in 5 mL LB at 37 °C to saturation during the day.
2. Streak on LB plates and grow overnight at 37 °C.
3. Pick single colonies and streak on LB_{AP} (negative control) and LB plates and grow overnight at 30 °C or 37 °C, respectively, to verify loss of the plasmid.

3.5 Generation of Multiple Deletion Strains Using Strain SIJ488 or Other Strains with the Recombineering Genes Integrated into the Genome (See Also Notes 4–6)

Prior to start it is important to design a strategy for which genes to be deleted together. It is recommended to select genes in relatively close proximity to each other when making two deletions in one round of electroporation. A circular map showing the distribution of genes to be deleted can, e.g., be made for *E. coli* K-12 MG1655 by using Ecogene Circle Maps, <http://www.ecogene.org/old/genemap/map.php> [23]. It is furthermore highly recommended to use primers with phosphorothioate bonds, and 100 bp homology extensions.

3.5.1 First and Second Gene Deletions

1. Pick a colony with the recombineering genes integrated into the genome (e.g., strain SIJ488) and grow in 250 mL shake flasks with shaking in 50 mL preheated LB until OD₆₀₀ ~0.2–0.3.
2. Add 750–1000 µL 1 M filter-sterilized arabinose (to induce the lambda Red recombineering genes).
3. Grow for 30–45 min.
4. Pour cells into 50 mL falcon tubes and place on ice for 5–10 min.
5. Prepare electrocompetent cells as described above.

6. Mix 2.5 μL of PCR product 1 (gene deletion 1 cassette) and 2.5 μL of PCR product 2 (gene deletion 2 cassette).
7. Add the 5 μL PCR mixture to the middle of the cell pellet and stir gently a couple of times with the pipet tip, do not pipet up and down. Keep the cells cold.
8. Transfer electrocompetent cells with PCR product to pre-cooled electroporation cuvettes.
9. Electroporate at 1.8 kV \sim 5 ms.
10. Add 1 mL LB **immediately** after electroporation and transfer back to 1.5 mL Eppendorf tube.
11. Grow cells with shaking at 37 $^{\circ}\text{C}$ for 1.5–2 h.
12. Plate on appropriate selection plate (e.g., LB_{Km} + C_m plates) and grow overnight at 37 $^{\circ}\text{C}$.

3.5.2 Removal of Antibiotic Cassettes and Third Gene Deletion

1. Transfer \sim 2–8 colonies to, e.g., PCR tubes with 20 μL LB.
2. Use 1–2 μL to perform colony PCRs using appropriate primers.
3. While colony PCR is running, grow cell cultures in 1 mL LB w. appropriate antibiotics (e.g., Km + Cm) at 37 $^{\circ}\text{C}$ with shaking.
4. After 2–3 h growth, spin down a verified pre-grown culture at 6500 $\times g$ for 2 min. Remove the supernatant and dissolve pellet in 1 mL LB + 50 mM rhamnose (start OD \sim 0.1–0.2).
5. Grow with induction for 4 h, spin down and transfer cells to 25–50 mL LB.
6. Continue as for gene deletion 1–2. It is recommended to only delete one gene after the removal of two antibiotic cassettes, while two genes can be deleted after the removal of only one antibiotic cassette. Note that an alternative antibiotic cassette than the ones just removed needs to be used during this step.

3.5.3 Verification of Removed Antibiotic Cassettes and Further Gene Deletions

1. Transfer \sim 2–8 colonies to, e.g., individual PCR tubes with 20 μL LB.
2. Do colony PCRs on all modified targets using appropriate upstream and downstream primers.
3. While colony PCR is running, transfer colonies to 1 mL LB + appropriate antibiotic(s) and grow at 37 $^{\circ}\text{C}$.
4. Transfer verified clones and continue as described above (from Subheading 3.5.2, **step 4**).

3.5.4 Removal of Integrated Recombineering Genes

1. Prepare two PCR products using the following primers and targeting two different antibiotics cassettes (note the 50 bp extensions are applicable for strain SIJ488).

- (a) SIJrem-1fw
ccgacgttgaccagccgcgtaacctggcaaaatcggttacggttgagtaaGTG
GTAGGCTGGAGCTGCTTC
- (b) SIJrem-1rv (targeting pKD4)
tttcggataacagaaaggccgggaataaccagcctcgctttgtaacggaAGAG
CGCTTTTGAAGCTG
- (c) SIJrem-2fw
agcttgcttcgtaattattggggaccctggattctcaccaataaaaaacGAACT
TCGGAATAGGAACTTCA
- (d) SIJrem-2rv (similar to SIJ4881&2rv)
aaaacataacaggaagaaaaatgcccgcttacgcaggcatccatttaCAT
ATGAATATCCTCCTTAGTTCC
2. Prepare electrocompetent cells as described above using the PCR mixture described in Subheading 3.5.4, **step 1** in the electroporation step.
 3. Plate on dual antibiotic plates and grow overnight at 37 °C.
 4. Verify insertions by performing colony PCRs.
 5. While colony PCRs are running grow the strains in LB + antibiotics for 2–3 h.
 6. Spin down a correct culture and induce the flippase recombinase as described above.
 7. Streak on LB plates and grow overnight at 37 °C.
 8. Verify removal of the integrated recombineering systems as well as the antibiotic cassettes by colony PCR.

4 Notes

1. Saving glycerol stocks with pSIJ8 can potentially cause curing problems. We have successfully frozen *E. coli* K-12 MG1655 and *E. coli* W strains with the plasmid without having any problems with curing later on. However, for *E. coli* BL21 this issue has been observed.
2. The colonies after overnight growth on plates can be small when using pSIJ8, and proper sized colonies may need longer incubation times than expected.
3. Problems in obtaining recombinant strains have occurred when plated on premade plates containing 100 µg/mL Ap and e.g., 50 µg/mL Km. It is therefore recommended to just spread 75 µL LB_{Ap} on an appropriate antibiotic plate.
4. Prior to freezing glycerol stocks it is important to restreak the cultures to ensure that a single homogenous population is saved.

5. During the induction of the flippase recombinase, old scars can cause unwanted recombineering events. It is therefore necessary to perform colony PCRs on all genes deleted after verification of the excision of the integrated antibiotic cassettes to ensure that such genomic re-arrangements have not occurred. The short induction time of the flippase recombinase limits the possibility for re-arrangements; however, they can still occur.
6. Regrowing the strains in the presence of the antibiotics applicable prior to the removal of the cassettes is important due to the presence of non-resistant, non-growing, but viable satellite colonies on the plate.

Acknowledgments

The development of the methods described in this chapter was obtained through funding from the Novo Nordisk Foundation. pKD3, pKD4, pSIJ8, and strain SIJ488 are all available from Addgene. Other plasmids are available upon request from the authors.

References

1. Karu AE, Sakaki Y, Echols H, Linn S (1975) The gamma protein specified by bacteriophage gamma. Structure and inhibitory activity for the recBC enzyme of *Escherichia coli*. *J Biol Chem* 250(18):7377–7387
2. Murphy KC (1991) Lambda gam protein inhibits the helicase and chi-stimulated recombination activities of *Escherichia coli* RecBCD enzyme. *J Bacteriol* 173(18):5808–5821
3. Kulkarni SK, Stahl FW (1989) Interaction between the sbcC gene of *Escherichia coli* and the gam gene of phage lambda. *Genetics* 123(2):249–253
4. Subramanian K, Rutvisuttinunt W, Scott W, Myers RS (2003) The enzymatic basis of processivity in lambda exonuclease. *Nucleic Acids Res* 31(6):1585–1596
5. Little JW (1967) An exonuclease induced by bacteriophage lambda. II Nature of the enzymatic reaction. *J Biol Chem* 242(4):679–686
6. Karakousis G, Ye N, Li Z, Chiu SK, Reddy G, Radding CM (1998) The beta protein of phage lambda binds preferentially to an intermediate in DNA renaturation. *J Mol Biol* 276(4):721–731. doi:10.1006/jmbi.1997.1573
7. Kmiec E, Holloman WK (1981) Beta protein of bacteriophage lambda promotes renaturation of DNA. *J Biol Chem* 256(24):12636–12639
8. Radding CM, Rosensweig J, Richards F, Casuto E (1971) Separation and characterization of exonuclease B protein, and a complex of both. *J Biol Chem* 246:2510–2512
9. Yu D, Ellis HM, Lee EC, Jenkins NA, Copeland NG, Court DL (2000) An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proc Natl Acad Sci U S A* 97(11):5978–5983. doi:10.1073/pnas.100127597
10. Yu D, Sawitzke JA, Ellis H, Court DL (2003) Recombineering with overlapping single-stranded DNA oligonucleotides: testing a recombination intermediate. *Proc Natl Acad Sci U S A* 100(12):7207–7212. doi:10.1073/pnas.1232375100
11. Sharan SK, Thomason LC, Kuznetsov SG, Court DL (2009) Recombineering: a homologous recombination-based method of genetic engineering. *Nat Protoc* 4(2):206–223. doi:10.1038/nprot.2008.227
12. Maresca M, Erler A, Fu J, Friedrich A, Zhang Y, Stewart AF (2010) Single-stranded heteroduplex intermediates in lambda Red homologous recombination. *BMC Mol Biol* 11:54. doi:10.1186/1471-2199-11-54
13. Mosberg JA, Lajoie MJ, Church GM (2010) Lambda Red recombineering in *Escherichia coli* occurs through a fully single-stranded

- intermediate. *Genetics* 186(3):791–799. doi:[10.1534/genetics.110.120782](https://doi.org/10.1534/genetics.110.120782)
14. Jensen SI, Lennen RM, Herrgard MJ, Nielsen AT (2015) Seven gene deletions in seven days: fast generation of *Escherichia coli* strains tolerant to acetate and osmotic stress. *Sci Rep* 5:17874. doi:[10.1038/srep17874](https://doi.org/10.1038/srep17874)
 15. Cherepanov PP, Wackernagel W (1995) Gene disruption in *Escherichia coli*: TcR and KmR cassettes with the option of FLP-catalyzed excision of the antibiotic-resistance determinant. *Gene* 158(1):9–14
 16. Datsenko KA, Wanner BL (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci U S A* 97(12):6640–6645. doi:[10.1073/pnas.120163297](https://doi.org/10.1073/pnas.120163297)
 17. Cox MM (1983) The FLP protein of the yeast 2-microns plasmid: expression of a eukaryotic genetic recombination system in *Escherichia coli*. *Proc Natl Acad Sci U S A* 80(14):4223–4227
 18. Gronostajski RM, Sadowski PD (1985) The FLP protein of the 2-micron plasmid of yeast. Inter- and intramolecular reactions. *J Biol Chem* 260(22):12328–12335
 19. Senecoff JF, Bruckner RC, Cox MM (1985) The FLP recombinase of the yeast 2-micron plasmid: characterization of its recombination site. *Proc Natl Acad Sci U S A* 82(21):7270–7274
 20. Yamamoto N, Nakahigashi K, Nakamichi T, Yoshino M, Takai Y, Touda Y, Furubayashi A, Kinjyo S, Dose H, Hasegawa M, Datsenko KA, Nakayashiki T, Tomita M, Wanner BL, Mori H (2009) Update on the Keio collection of *Escherichia coli* single-gene deletion mutants. *Mol Syst Biol* 5:335. doi:[10.1038/msb.2009.92](https://doi.org/10.1038/msb.2009.92)
 21. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2(2006):0008. doi:[10.1038/msb4100050](https://doi.org/10.1038/msb4100050)
 22. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1):27–30
 23. Zhou J, Rudd KE (2013) EcoGene 3.0. *Nucleic Acids Res* 41(Database issue):D613–D624. doi:[10.1093/nar/gks1235](https://doi.org/10.1093/nar/gks1235)

Designing and Implementing Algorithmic DNA Assembly Pipelines for Multi-Gene Systems

Szu-Yi Hsu and Michael J. Smanski

Abstract

Advances in DNA synthesis and assembly technology allow for the high-throughput fabrication of hundreds to thousands of multi-part genetic constructs in a short time. This allows for rapid hypothesis-testing and genetic optimization in multi-gene biological systems. Here, we discuss key considerations to design and implement an algorithmic DNA assembly pipeline that provides the freedom to change nearly any design variable in a multi-gene system. In addition to considerations for pipeline design, we describe protocols for three useful molecular biology techniques in plasmid construction.

Key words Golden Gate assembly, Isothermal assembly, PCR-ligation, DNA assembly, Genetic refactoring, BioDesign automation

1 Introduction

In the past 50 years, recombinant DNA (rDNA) technology has been applied to produce life-saving medicines, herbicide-resistant or nutrient-enriched food crops, and has enabled to numerous discoveries in the life sciences. The vast majority of rDNA-enabled applications rely on rather simple systems comprising one or a few genes. Recent advances in DNA synthesis, DNA assembly, and synthetic biology have allowed for the engineering of more complex biological capabilities that require the coordinated expression of a dozen or more genes. These include multi-gene systems that control cellular computation [1, 2], biosynthesis of structurally complicated drugs [3, 4] and enhanced biomaterials [5], and energy systems [6]. The engineering of these massively multi-part genetic systems is facilitated by integrated pipelines of automated genetic design, plasmid construction, and experimental characterization. Engineering biology at this scale is facilitated by robust and algorithmic DNA assembly pipelines that allow hundreds of variant multi-gene constructs to be built and tested in parallel.

DNA synthesis and DNA assembly are two related yet distinct technology areas that both have seen tremendous advances in the past decade. We define DNA synthesis as the process of polymerizing nucleotide monomers using chemical processes. This is most commonly done with phosphoramidite chemistry, but the platforms vary from solid phase synthesis on resin to photochemistry on glass slides. DNA synthesis technology has been reviewed recently [7] and is not the focus of this chapter. DNA assembly refers to the suite of biochemical methods available to combine individual fragments of single- or double-stranded DNA into larger composite constructs. DNA assembly has been possible for the past half century using restriction enzyme-mediated cloning reactions; however, a number of new techniques now allow DNA assembly of many parts with high levels of efficiency and complete control over genetic design. For genetic engineering projects that require construction and testing of combinatorial libraries of constructs composed of the same DNA sequences (e.g., genes in a metabolic pathway), DNA assembly is more economical than DNA synthesis alone. A DNA assembly “pipeline” is a set of protocols that can be used in succession to go all the way from the individual fragments of synthetic DNA to a large multi-gene construct.

Refactored systems, in which the genetics have been systematically rewritten to enable engineering efforts, can be designed to be compatible with any DNA assembly pipeline. The main advantages to working with refactored genetic systems are that (1) every functional genetic element is characterized and its role in the expression of the system is understood, (2) regulation can be decoupled from host chromosome, for example by using orthogonal RNA polymerases, and (3) they contain a modular genetic architecture. This allows the “parts approach” in synthetic biology, where genetic elements like promoters and ribosome binding sites are swapped in a combinatorial fashion to control gene expression. Many complex systems have been refactored in recent years, including bacteriophage [8], nitrogen fixation [9], and natural product biosynthesis pathways [10, 11]. Even a 272 kb yeast chromosome has been redesigned and assembled from synthetic oligonucleotides [12].

An important facet of refactored systems is that they can be designed to accommodate the DNA synthesis and assembly protocols used for fabrication. This is in contrast to modification of plasmids encoding wild-type sequences, which is highly constrained by the presence or absence of restriction recognition sequences and by the overlapping nature of encoded genetic elements. This means that unique DNA assembly routes had to be created for each recombinant plasmid design on a case-by-case basis. With refactored systems, it is possible to predefine a set of robust DNA assembly protocols that will be used repeatedly to build any construct via an algorithmic, efficient pipeline. In other words, the DNA sequence is adjusted to accommodate the DNA assembly,

not vice versa. Diversity in the genetic designs is controlled by varying the substrate DNA fragments that are included in each assembly reaction.

Recently, many novel assembly techniques have been developed for large multi-part assembly for multi-gene systems from small DNA fragments, including isothermal assembly [13], Golden Gate assembly [14], and yeast recombination [15]. These cloning methods can be coupled with standardized vector designs to provide hierarchical DNA assembly systems, such as Modular Cloning [16], GoldenBraid [17], BioBrick/Bglbrick [18], and TNT-cloning [19]. However, understanding the strengths and weaknesses of each cloning method allows the genetic engineering to customize a DNA assembly system to meet the design constraints of a specific multi-gene engineering project.

This review covers major design considerations of an algorithmic DNA assembly pipeline and three useful DNA assembly techniques that can be incorporated in the assembly pipeline that fit individual assembly projects. We describe protocols to design primers and perform PCR-ligation, isothermal assembly, and Golden Gate assembly, and provide an example of an algorithmic DNA assembly pipeline that integrates all the three techniques. Lastly, we will discuss the advantages and limitations of each assembly technique so that the users are aware of how to design, implement, test, and troubleshoot an algorithmic pipeline for given DNA assembly applications.

2 Materials

Prepare all the enzymatic reactions in PCR tubes and on ice. Prevent unnecessary thawing and freezing of the buffers and enzymes, which are generally heat labile and therefore should always be kept at -15 to -25 °C. When preparing reactions, the enzymes can be kept on ice or small freezer boxes for a brief amount of time. Always prepare reactions using ultrapure water (which is often purified by Milli-Q[®] water purification system) that has a resistivity of 18.2 M Ω cm. All the buffers and enzymatic reactions should be mixed well before use.

2.1 Computational and Instrumental Requirements

1. Plasmid editing software facilitates primer design and visualization of intermediate and final constructs. Free software that the authors recommend includes ApE (A Plasmid Editor) for computers running the Windows operating system. The software can be downloaded at <http://biologylabs.utah.edu/jorgensen/wayned/ape/>. For Mac users we recommend SnapGene (<http://www.snapgene.com/>).
2. Nanodrop or similar spectrophotometer.

3. Thermocycler.
4. DNA gel electrophoresis equipment.
5. Gel imaging system.

2.2 PCR-Ligation

1. NEB T4 DNA ligase (New England Biolabs Inc.).
2. NEB T4 DNA polynucleotide kinase (New England Biolabs Inc.).
3. NEB 10× T4 ligase buffer (New England Biolabs Inc.).
4. NEB DpnI (New England Biolabs Inc.).
5. Q5[®] High-Fidelity PCR Kit (New England Biolabs Inc.).
6. Deoxynucleotide (dNTP) Solution Mix.
7. Plasmid template(s).
8. Primers.
9. PCR tubes.
10. Zymoclean[™] Gel DNA Recovery Kit (Zymo Research).
11. Agarose gel for DNA gel electrophoresis.
12. 6× DNA loading Dye.
13. 1 kb DNA plus ladder (Thermo Fisher Scientific).

2.3 Isothermal Assembly

1. NEB DpnI (New England Biolabs Inc.).
2. Q5[®] High-Fidelity PCR Kit (New England Biolabs Inc.).
3. Deoxynucleotide (dNTP) Solution Mix.
4. NEB Gibson Assembly 2× Master mix (New England Biolabs Inc.).
5. Zymoclean[™] Gel DNA Recovery Kit (Zymo Research).
6. Up to eight DNA fragments with 20–40 bp overlapping sequences.
7. PCR tubes.
8. Plasmid template(s).
9. Primers.
10. PCR tubes.

2.4 Golden Gate Assembly

1. Type II restriction endonuclease.
2. Promega T4 DNA ligase (HC) (Promega).
3. 10× T4 DNA ligase buffer (Promega).
4. Plasmid DNA with Type II restriction recognition sites flanking the desired sequences and compatible 5' overhangs.
5. Primers.
6. PCR tubes.

3 Methods

3.1 Key Considerations for Designing an Algorithmic DNA Assembly Pipeline

Developing a customized DNA assembly pipeline can be organized into three stages. The first stage is definition of design parameters for genetic constructs needed for the research project (Fig. 1). For example, what types of genetic elements (promoters, ribosomal binding sites (RBSs), insulator sequences, etc.) will be varied during library design? Where are scar sequences (vestigial DNA sequences that remain as a result of the cloning strategy, for example, restriction recognition sequences) between genetic elements to be tolerated, and how large can these scars be? Will higher-level

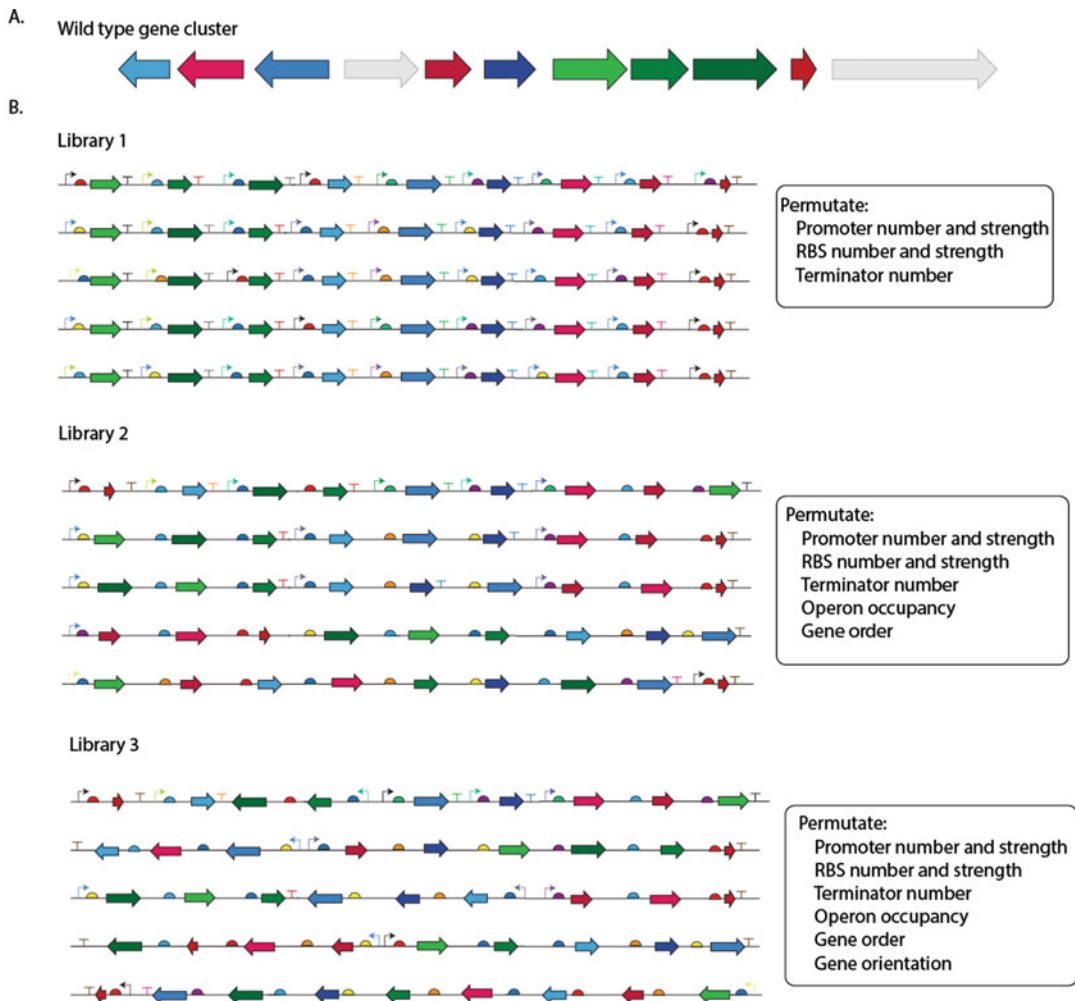


Fig. 1 Three library designs with different design parameters. (a) A wild-type gene cluster. *Arrows* represent genes. *Gray arrows* are the genes not essential to gene functions. (b) Example of three libraries with different design parameters

architectural variables including operon occupancy, gene order, or gene orientation be permuted? Libraries that contain more design constraints can often be constructed with more streamlined assembly pipelines compared with libraries that encode diverse designs. In the second stage, the DNA assembly methods most suited to the assembly needs are selected. In the Notes section below, we discuss the advantages, limitations, failure modes, and troubleshooting methods for three DNA assembly techniques: PCR-ligation, isothermal assembly, and Golden Gate assembly. Other assembly techniques have been reviewed recently [20]. Included in the second stage is specifying the specific molecular components that will be used for DNA assembly (*see Note 1*). This is particularly important for Golden Gate assembly, where there are many type II restriction enzymes that could be used in a pipeline, the selection of which will impact genetic part domestication. Aside from restriction enzyme choice, other components to determine include molecular features of cloning and expression vectors, antibiotic resistance cassette as selectable markers, and reporter genes for rapid and efficient screening of correct construct in each step of the pipeline. With these variables set, it is possible to construct all required vector plasmids needed for the assembly system. The third stage is to plan how intermediate and final constructs will be validated. Plasmid verification can take even more time and resources than the DNA assembly, and the unique failure modes of each technique dictate different verification methods (*see Note 2*). The following subsections cover three useful DNA assembly methods that can be integrated into an algorithmic pipeline. Special considerations concerning the strengths and weaknesses of each method (*see Note 3*), and an example hierarchical DNA assembly pipeline (*see Note 4*, Fig. 2) are included in the Notes section below.

3.2 PCR-Ligation

PCR-ligation is useful for adding or modifying short DNA sequences (1–200 bp) in existing plasmids. The entire plasmid is amplified in a single PCR reaction, with new sequences added via the 5'-ends of the forward and/or reverse primers (Fig. 3a). Intramolecular blunt-end ligation of the linear PCR product re-circularizes the plasmid and readies it for transformation. PCR-ligation is particularly useful to build libraries of small functional DNA sequences such as promoters, RBS, and terminators with flanking sequences that will allow them to enter later stages of an assembly pipeline.

3.2.1 Primer Design

1. The 3'-ends of both the forward and reverse primers should have sufficient complementarity to the template to provide a T_m of 55–70 °C (usually 18–35 bp that will hybridize to the template; this can be checked using New England Biolabs' T_m calculator, <http://tmcalculator.neb.com>). Note that the primers should anneal at the regions of the template plasmid

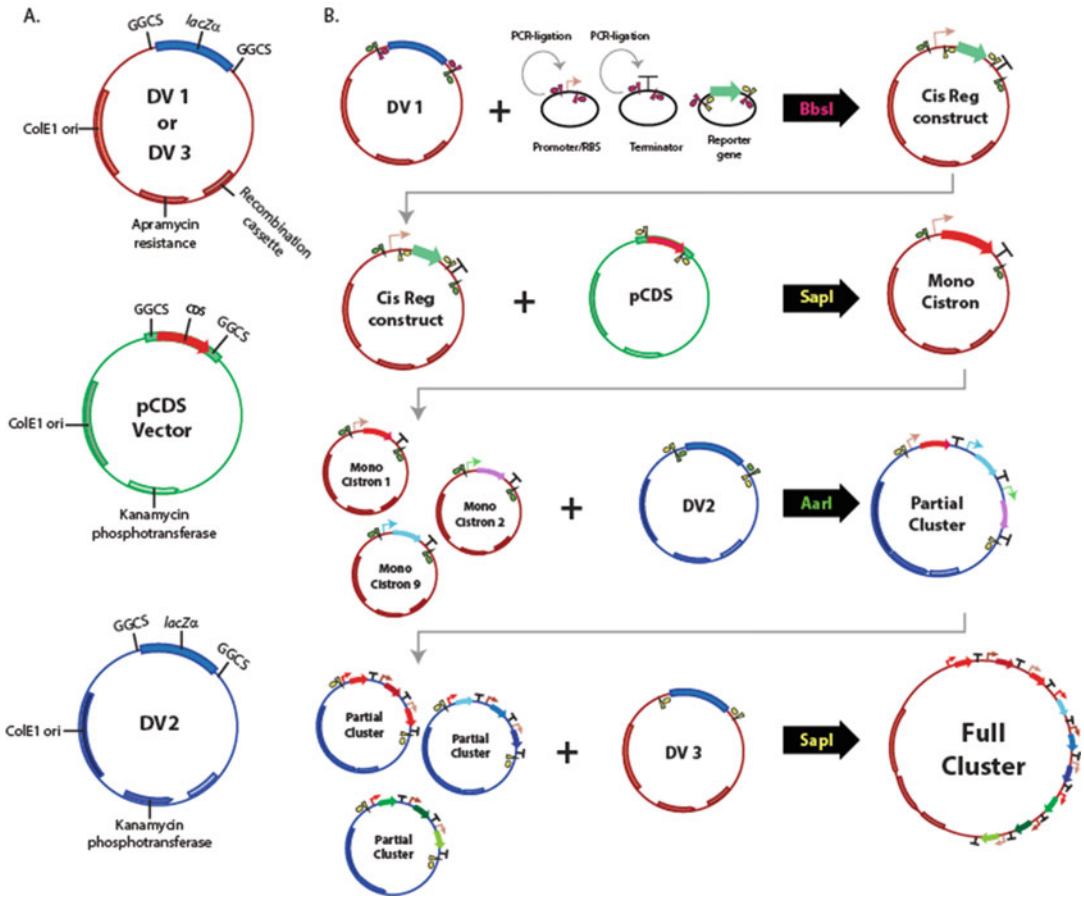


Fig. 2 Example of an algorithmic DNA assembly pipeline incorporating isothermal assembly, PCR-ligation, and Golden Gate assembly. *Green arrow* represents reporter gene. *GGCS* Golden Gate Cassette Site, *DV* destination vector. The *black arrow* contains the name of type II restriction enzyme used in individual level of Golden Gate assembly

that allow the user to amplify molecular features that the user wants to include in the final construct.

2. New DNA sequence is added to the 5' end of either forward primer, reverse primer or both and will be incorporated at the ligation junction. It is highly suggested to use plasmid-editing software to construct the desired final sequences in silico before designing primers.

3.2.2 Protocol

1. Thaw dNTPs, Q5 PCR buffer, primers, and template DNA on ice.
2. Prepare the PCR mix on ice according to the standard NEB Q5[®] polymerase PCR protocol (Table 1). Add Q5 polymerase to mix the last and avoid letting polymerase warm above -20 °C.

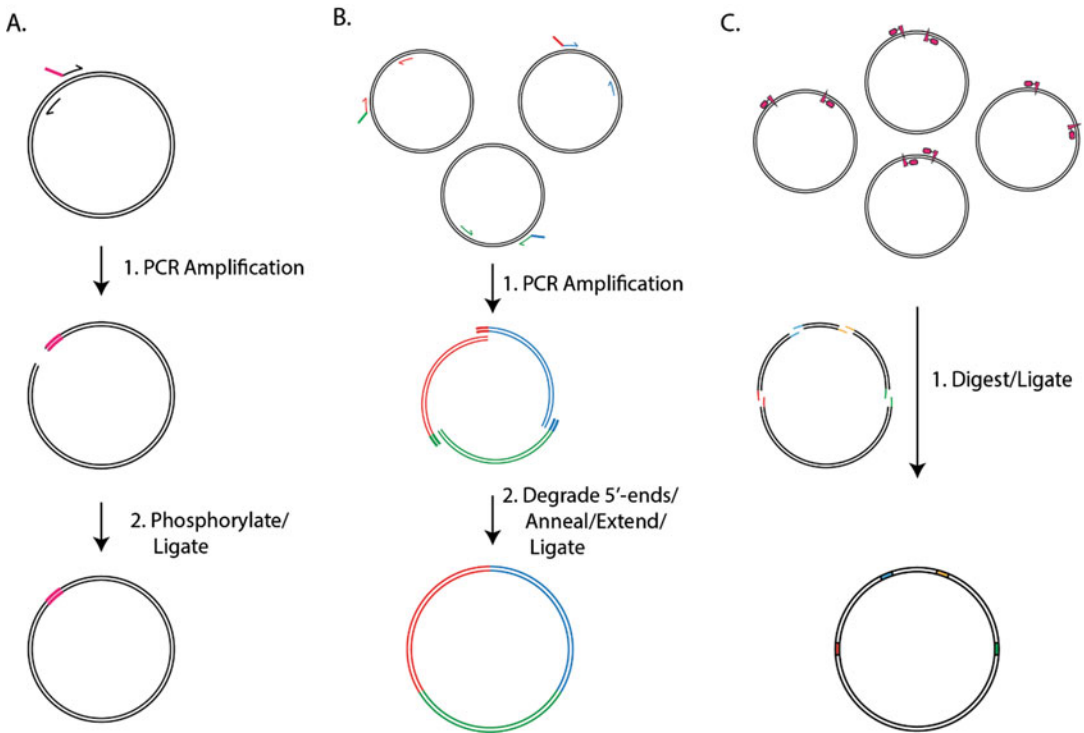


Fig. 3 Overall schematics of PCR-ligation, isothermal assembly, and Golden Gate assembly. **(a)** Key steps for PCR-ligation. *Half arrows* represent oligonucleotide primers. *Red color* denotes new DNA sequence being added to the final construct (*bottom*). **(b)** Isothermal assembly. The double-stranded lines with different colors denote PCR products with unique set of oligonucleotide primers. **(c)** Golden Gate assembly. *Red rectangle* represents arbitrary type IIS restriction recognition sequences. *Red triangle* represents arbitrary cleavage site. The single line highlighted in different colors represents arbitrary 5' overhang generated by restriction digest

3. Calculate the appropriate annealing temperature for specific primer pairs using the NEB T_m calculator (<http://tcalculator.neb.com/#!/>). Make sure to select “Q5” under product group drop down menu, “Q5 High-Fidelity DNA Polymerase” under polymerase/Kit drop down menu. Type “500” in Primer concentration (nM) box. Only input DNA sequences that anneal to the template. Otherwise, the calculated anneal temperature will be overestimated if input is the entire primer sequence.
4. Calculate the extension time for your construct. Q5 DNA polymerase extends at 20–30 s per kilobase.
5. Start the thermal cycler reaction with the following parameters given in Table 2.
6. Once the PCR is finished, freeze the mixture in the $-20\text{ }^{\circ}\text{C}$ or proceed to **step 7**.

Table 1
Composition of Q5 PCR reaction mixture

| Component | Volume | Final concentration |
|-----------------------------------|----------|---------------------|
| 5× Q5 reaction buffer | 5 μL | 1× |
| 10 mM dNTPs | 0.5 μL | 200 μM |
| 10 μM forward primer | 1.25 μL | 0.5 μM |
| 10 μM reverse primer | 1.25 μL | 0.5 μM |
| Template DNA | Variable | <1000 ng |
| Q5 high-fidelity DNA polymerase | 0.25 μL | 0.02 U/μL |
| 5× Q5 high GC enhancer (optional) | (5 μL) | (1×) |
| Nuclease-free water | To 25 μL | |

Table 2
Thermocycling program for Q5 reaction

| Step | Temperature | Time |
|----------------------|-------------|------------|
| Initial denaturation | 98 °C | 30 s |
| 25–35 cycles | 98 °C | 5–10 s |
| | *50–72 °C | 10–30 s |
| | 72 °C | 20–30 s/kb |
| Final extension | 72 °C | 2 min |
| Hold | 4–10 °C | ∞ |

*Annealing temperature depends on primer sequence

7. Analyze the presence of the PCR product by DNA gel electrophoresis: mix 4 μL PCR and with 0.6 μL 6× DNA loading dye. Load the PCR mixture and 1 kb DNA ladder on a 1% agarose TAE gel to check for expected product size.
8. *DpnI* digest to remove template plasmid: add 1 μL *DpnI* to the remaining PCR mixture and incubate at 37 °C for 30 min to 1 h. Heat inactivate *DpnI* at 80 °C for 20 min.
9. Purify the digested mixture using Zymoclean Gel Extraction Kit: Add 5 volume of DNA binding buffer to your PCR and load to the column. Spin down for 30 s. Discard flow-through. Wash each column using 200 μL column wash buffer two times. Add minimal amount of water (≥ 6.5 μL) to elute the purified PCR product, so it is as concentrated as possible.
10. Perform T4 DNA ligation reaction: For each ligation, add the entire purified PCR product, 1 μL T4 ligase, 5 units T4

polynucleotide kinase (PNK), 1 μL T4 ligase buffer, and add water to bring up the final reaction volume to 10 μL . T4 ligase and T4 PNK should be added to the reaction last. Incubate the ligation mixture at room temperature for 1 hour. Transform 1–5 μL of the ligation mixture into 20–50 μL chemically competent cells.

3.3 Isothermal Assembly

Isothermal assembly [13], also known as Gibson assembly, is a homology-based, restriction enzyme-independent method for stitching together multiple pieces of linear DNA (Fig. 3b). The reaction utilizes a master mix of a 5' exonuclease to “chew back” one strand of the double-stranded DNA, a DNA polymerase that fills in the gaps that are created and eventually overtake the exonuclease, and a DNA ligase that covalently joins independent pieces by repairing nicks. This method is fast, efficient, and reliable for multipart DNA assembly reactions, and has been used in the complete chemical synthesis of a bacterial genome [21].

3.3.1 Primer Design

1. Primers for adjoining DNA fragments must encode 20–40 bases of overlapping sequence for annealing and ligation. We highly recommend designing the final construct sequences in silico prior to primer design. For each DNA fragment to be assembled, first design annealing 3'-ends of both the forward and reverse primers. They should have sufficient complementarity to the template to provide a T_m of 55–70 $^{\circ}\text{C}$ (usually 18–35 bp that will anneal to the template).
2. Overlap sequences of 20–40 nucleotides from neighboring DNA fragments are added into the 5' end of both the forward and reverse primers to allow for annealing of overhangs.
3. For sections of DNA that must be completely synthesized de novo, single-stranded oligos can be used directly by “tiling” them with successive 20 bp overlaps. The key to this single-stranded Gibson assembly is that the terminal oligo on each end must create a 3' overhang.
4. Alternatively, the primers can be designed by NEBuilder, a web-based Gibson assembly visualization tool provided by NEB. This can be found at <http://nebuilder.neb.com>.

3.3.2 Protocol

1. Perform PCR with primers designed to incorporate overlapping ends as described in the PCR protocol in Subheading 3.2.2, including *DpnI* digestion and purification.
2. Measure the concentration of purified DNA fragments using Nanodrop.
3. Thaw 10 μL Gibson assembly master mix on ice.
4. Add 100 ng of the vector and equimolar amounts of other DNA, and add water to bring the final volume up to 20 μL .

5. Incubate the isothermal reaction mixture at 50 °C for 1 h.
6. Transform 1–5 μL of the isothermal reaction into 20–50 μL chemically competent cells.

3.4 Golden Gate Assembly

Golden Gate assembly [14] utilize type II restriction enzymes to generate 3–4 bp sticky ends outside their recognition sequences that can be subsequently joined by T4 ligase (Fig. 3c) in a one-pot reaction. That Type II restriction enzymes cleave outside their recognition site provides several advantages. First, the overhangs generated upon cleavage can be customized because they are independent of the restriction recognition sequences, allowing for scarless assemblies or combinatorial assembly between user-defined 3–4 base junctions. Also, because recognition sites are not present in the final assembled product, Golden Gate assembly can proceed in a one-pot digestion/ligation reaction with substantially higher efficiency than traditional cloning. Golden Gate reactions can routinely be used to incorporate several (>5) fragments into large (>25 kb) plasmid designs. An advanced version of Golden Gate assembly utilizes two type II restriction enzymes alternating between assembly stages to build an “infinite cloning loop” (Fig. 4).

3.4.1 Primer Design

As with Isothermal assembly reactions, we suggest creating a plasmid sequence file of the final construct before designing primers. Here, we discuss two aspects of Golden Gate assembly that require extra attention: the orientation of restriction recognition sequence and the position and design of scar site. Note that unlike isothermal reactions, it is possible and even preferable to start with circular plasmid substrates instead of linear fragments. Whether using circular or linear substrates, it is essential to design the orientation of restriction recognition sites and overhangs correctly.

1. For each DNA fragment, design the annealing 3' ends of both the forward and reverse primers. They should have sufficient complementarity to the template to provide a T_m of 55–70 °C (usually 18–35 bp that will anneal to the template).
2. Orientation of the restriction recognition site: Type II restriction recognition sites are not palindromic, and DNA cleavage occurs on one side of the site (represented by carrot symbols in Fig. 3c). It is important to orient each restriction recognition site so that cleavage occurs between the recognition site and the fragment to be assembled. Cleavage with the Type II restriction enzymes will produce an insert that lacks the recognition site.
3. Design of annealing cohesive ends: The spacing between restriction recognition sequence and cleavage site depends on the choice of type II enzyme. Similar cohesive ends should be avoided in a single Golden Gate reaction, and the likelihood of two cohesive ends to join during the digestion/ligation reaction can be predicted using thermodynamic models [22].

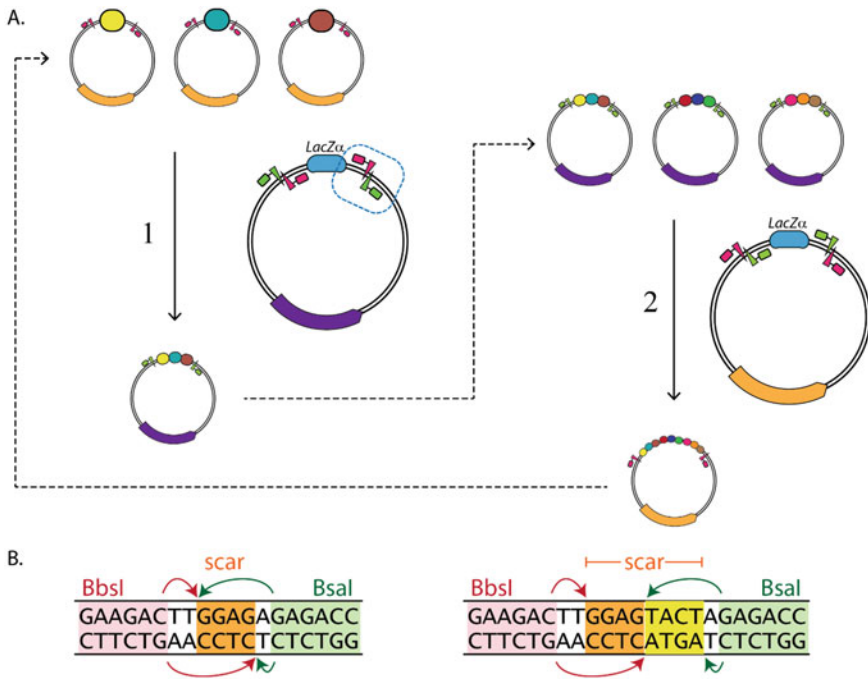


Fig. 4 Schematic of an iterative Golden Gate assembly. **(a)** Alternating between two arbitrary type IIs restriction enzyme allows infinite cloning loop. The oval in different colors denotes DNA part. The arbitrary antibiotic resistance cassette on the donor and destination vector are highlighted in *orange* and *purple*. Type IIs restriction recognition sequence is represented as *rectangle*. The cleavage site is represented as *triangle*. **(b)** Two design strategies of customizable scar sites. The *left diagram* shows that the directionality of *BbsI* and *BsaI* recognition sequence embedded in the plasmid shares the same cut site. The *right diagram* shows different scar sites generated by *BsaI* or *BbsI* site. The cohesive ends can be customized based on how the recognition sites are positioned relative to each other and the Type IIs recognition sequence. *Pink box*: *BbsI* recognition sequence. *Green box*: *BsaI* recognition sequence. *Orange and yellow box*: scar sequence generated by either *BbsI* or *BsaI* cleavage

3.4.2 Protocol

1. Measure the DNA concentration of each fragment to be assembled on a NanoDrop spectrophotometer.
2. Calculate the mass of each DNA fragment equivalent to 20 fmol: dilute the DNA stock so each DNA fragment is approximately 20 fmol/ μ L.
3. Prepare a 10 μ L reaction mixture by adding the DNA fragments, water, and 1 μ L 10 \times T4 ligase buffer. Finally, add type IIs restriction enzyme (10 U) and T4 ligase (10 U). If performing many reactions in parallel, the water, buffer, and enzymes can be combined to form a master mix immediately before adding the DNA fragments.
4. Mix the reaction by pipetting up and down three to four times and centrifuge briefly.

Table 3
Golden Gate assembly cycling conditions

| Step | Temperature | Time |
|--------------------------------------|--------------------|------------------------|
| Initial digest | 37 °C | 5 min |
| Initial ligation | 16 °C | 5 min |
| 10–30 cycles | 37 °C | 1 min |
| | 16 °C | 1 min |
| Ligase heat inactivation | 50 °C | 5 min |
| Restriction enzyme heat inactivation | 65 °C ^a | 10–20 min ^a |
| Hold | 4–10 °C | ∞ |

^aRestriction enzymes have different heat inactivation temperatures. Refer to vendor information to set the proper inactivation time and temperature

5. Incubate the reaction mixture on a thermal cycler using the cycling conditions described in Table 3.
6. Transform 1–5 μL of each Golden Gate assembly reaction into 20–50 μL chemically competent cells.

4 Notes

1. Molecular specifications for algorithmic DNA assembly.

Choice of type II restriction endonuclease. Many type II restriction endonucleases can be used in an algorithmic DNA assembly pipeline, including *BsaI*, *BbsI*, *AarI*, and *SapI*. Selection of suitable restriction endonuclease is influenced by (1) differences in cohesive end size (i.e., 3-base overhangs generated by enzymes like *EarI* vs 4-base overhangs generated by enzymes like *BsmBI*) and (2) frequency of recognition site in vectors or parts. Smaller cohesive ends generate smaller vestigial scars, but also decrease the number of fragments that can be reasonably assembled in a single reaction. More commonly, selection of restriction endonuclease is based on the frequency of the recognition sites in the substrate genetic parts or assembly vectors. These sites will have to be removed via synonymous mutations to “domesticate” genetic parts that will enter the assembly pipeline. For example, *AarI* recognition sites are less frequent than *BbsI* recognition sites in *Streptomyces* genes, so using *AarI* would be preferred for building multigene pathways from *Streptomyces* genes.

Choice of selectable marker in vector backbones. The vector backbone for initial, intermediate, and final constructs should be tailored for the specific assembly project. One failure mode for every DNA assembly method is the growth of colonies

Table 4
Failure modes and screening methods for assembly methods described here

| Assembly techniques | Failure modes | Best screening method |
|----------------------|---|---|
| PCR-ligation | Small deletions at site of ligation Point mutations in the construct introduced through PCR | DNA sequencing |
| Isothermal assembly | Point mutations in the construct introduced through PCR or junction gap-filling Low efficiency due to too many fragments, high GC fragments, or extremely long fragments | DNA sequencing |
| Golden Gate assembly | Missing fragments | Colony PCR or diagnostic restriction digest |

resulting from undigested substrate plasmids. This is easily avoided by changing selection markers in successive “levels” in a DNA assembly pipeline. For example, “monocistron” plasmids (Fig. 2) contain a kanamycin resistance marker while “partial cluster” plasmids contain an ampicillin marker. Using two unique selectable markers is sufficient for creating an infinite cloning loop (Fig. 4). Undigested destination vectors can be easily identified if a *lacZ* α reporter cassette is included between the Golden Gate cloning scars (Figs. 2 and 4).

Choice of origin of replication in vector backbones. High-copy origins of replication provide the best plasmid yields following purification, and thus are suggested for all vectors holding intermediate assembly constructs (i.e., partial clusters). However, the use of low- or medium-copy origins of replication might mitigate potential toxicity of constructs. The final destination vector should be customized for the expression host. Origin of replication copy number as well as the choice between replicative or chromosome-integrating vectors will depend on the specific project requirements.

2. Failure mode, screening techniques, and troubleshooting.

When screening plasmids to verify the fidelity of DNA assembly reactions, it is good to consider the common failure modes of individual methods. These are summarized in Table 4, where we consider Golden Gate cloning from previously sequence-verified plasmid substrates, not PCR products.

3. Strengths and weaknesses of assembly techniques.

PCR-Ligation. The advantage of PCR-ligation is its utility to generate large combinatorial library with relatively small amount of primers. It is particularly suited to building out large libraries of promoter-RBS pairs in a combinatorial fashion. Primer requirement scales with the sum of parts, not the product of parts, so assembling ten promoter sequences and

ten RBS sequences into 100 combinations only requires purchasing 20 oligonucleotides. Also, the use of blunt-end ligations yields scarless part-junctions, which can be useful between regulatory sequences in the 5'-UTR that are sensitive to relative spacing. Lastly, this is a restriction enzyme-independent technique that limits sequence constraints. The largest weakness of PCR-ligation is that it is limited to small genetic parts that can be fully encoded in an oligonucleotide primer. Another weakness is the most common failure mode of small deletions at the ligation junction, which require sequence verification and add to the validation costs.

Isothermal Assembly. Isothermal assembly is an incredibly useful method with diverse applications. Primer design is straightforward, and there are few sequence-constraints since it is restriction enzyme-independent. The ease of building scarless constructs allows the genetic engineer to specify every base in a plasmid design. Further, many fragments can be assembled in a single reaction with high efficiency. Drawbacks to isothermal assembly are that it does not lend itself to combinatorial assembly. Joining parts need ~20 bases of homologous sequence for the reaction to proceed meaning that for combinatorial libraries, either (1) a 20 base scar needs to be designed between neighboring genetic parts, or (2) primers need to be designed to accommodate every possible unique part junction to avoid scars. This latter option would make oligonucleotide requirements scale by the product of parts (i.e., a 10 promoter \times 10 RBS library would require ~100 oligonucleotides). Isothermal reactions aimed to produce constructs greater than ~12 kb require many-fold more substrate DNA, so this technique does not lend itself to large construct assembly. Lastly, secondary structure in the substrate fragments, for example that caused by high-GC DNA, can reduce the efficiency of this technique.

Golden Gate Assembly. Golden Gate assembly is known for its scalability and high efficiency. First, the ability to define annealing cohesive ends allows either scarless assembly or combinatorial assembly with small scar sequences. This method can readily be used to assemble >10 fragments in a single reaction, although reaction efficiencies decrease with more parts. Of the methods described here, Golden Gate assembly is the best choice for assembling large (>10 kb) constructs. Weaknesses include additional complexity in primer design compared with the other methods that requires substantial practice to become experienced. Also, as a restriction enzyme-dependent technique, substrate parts and vectors need to be domesticated prior to entry into the pipeline. However, this can be leveraged as an advantage, where substrate fragments are domesticated and cloned into a plasmid before entry into the pipeline.

Sequencing of these substrate plasmids decreases future validation costs and starting with plasmid substrates increases the efficiency of the reaction.

4. Sample DNA assembly pipeline with integrated expression analysis.

An example DNA assembly pipeline designed to accommodate different promoter strengths, RBS strengths, gene order, and operon occupancy is described below and illustrated in Fig. 2.

Prior to executing the assembly pipeline, all required vector backbones are constructed using isothermal assembly. Libraries of promoter-RBS combinations and terminators are built using PCR-ligation. All coding sequences are domesticated in the pCDS vector via isothermal assembly or Golden Gate assembly. All vectors and part plasmids should be sequence verified before proceeding further.

The first step in executing the assembly pipeline is using a BbsI Golden Gate assembly reaction to combine a promoter-RBS, a reporter gene, and a terminator together into DV1 to produce a “CisReg construct.” CisReg constructs contain all of the *cis*-regulatory elements surrounding a CDS. The reporter gene serves two purposes. First, it is a placeholder gene with cloning sites that will eventually allow for its scarless replacement with any CDS from the pCDS vector. Second, the reporter allows for quantification of or expression strength from each unique combination of *cis*-regulatory elements present in the library. Appropriate reporter gene expression assays (fluorescence, colorimetric, etc.) are performed at this point.

Next, the reporter genes are swapped out with appropriate CDSs from the pCDS library using a scarless SapI Golden Gate reaction. The resulting monocistronic construct is termed a “MonoCistron” part. The assembly scars flanking each MonoCistron part determine the eventual order and orientation of the genes in a final construct. Multiple MonoCistron parts are combined using an AarI Golden Gate assembly to yield a partial cluster. Finally, partial clusters are pieced together by a SapI Golden Gate assembly to produce a final cluster. Theoretically, alternating the last two type II restriction endonucleases, SapI and AarI, can create an infinite cloning loop as described in Fig. 4.

Acknowledgment

S-Y. H. is supported by Biocatalysis Grant from the University of Minnesota BioTechnology Institute.

References

1. Regev A, Shapiro E (2003) Cells as computation. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 2602:1–3. doi:[10.1007/3-540-36481-1_1](https://doi.org/10.1007/3-540-36481-1_1)
2. Brophy JAN, Voigt CA (2014) Principles of genetic circuit design. *Nat Methods* 11 (5):508–520. doi:[10.1038/nmeth.2926](https://doi.org/10.1038/nmeth.2926)
3. Ro D, Paradise EM, Ouellet M et al (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* 440:3–6. doi:[10.1038/nature04640](https://doi.org/10.1038/nature04640)
4. Ajikumar PK, Xiao W-H, Tyo KEJ et al (2010) Isoprenoid pathway optimization for Taxol precursor overproduction in *Escherichia coli*. *Science* 330(6000):70–74. doi:[10.1126/science.1191652](https://doi.org/10.1126/science.1191652)
5. Scheibel T (2004) Spider silks: recombinant synthesis, assembly, spinning, and engineering of synthetic proteins. *Microb Cell Factories* 3 (1):14. doi:[10.1186/1475-2859-3-14](https://doi.org/10.1186/1475-2859-3-14).
6. Miyao M (2003) Molecular evolution and genetic engineering of C4 photosynthetic enzymes. *J Exp Bot* 54(381):179–189. doi:[10.1093/jxb/54.381.179](https://doi.org/10.1093/jxb/54.381.179).
7. Kosuri S, Church GM (2014) Large-scale de novo DNA synthesis: technologies and applications. *Nat Methods* 11(5):499–507. doi:[10.1038/nmeth.2918](https://doi.org/10.1038/nmeth.2918)
8. Chan LY, Kosuri S, Endy D (2005) Refactoring bacteriophage T7. *Mol Syst Biol* 1:2005.0018. doi:[10.1038/msb4100025](https://doi.org/10.1038/msb4100025)
9. Lauren C, Lauren BA, Smanski MJ, et al. Functional optimization of gene clusters by combinatorial design and assembly. 2014. doi:[10.1038/nbt.3063](https://doi.org/10.1038/nbt.3063).
10. Shao Z, Rao G, Li C, Abil Z, Luo Y, Zhao H (2013) Refactoring the silent spectinabilin gene cluster using a plug-and-play scaffold. *ACS Synth Biol* 2(11):662–669. doi:[10.1021/sb400058n](https://doi.org/10.1021/sb400058n)
11. Oßwald C, Zipf G, Schmidt G et al (2014) Modular construction of a functional artificial epothilone polyketide pathway. *ACS Synth Biol* 3:759–772. doi:[10.1021/sb300080t](https://doi.org/10.1021/sb300080t)
12. Annaluru N, Muller H, Mitchell LA et al (2014) Total synthesis of a functional designer eukaryotic chromosome. *Science* 344:55–59. doi:[10.1126/science.1249252](https://doi.org/10.1126/science.1249252)
13. Gibson DG, Young L, Chuang R-Y et al (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6 (5):343–345. doi:[10.1038/nmeth.1318](https://doi.org/10.1038/nmeth.1318)
14. Engler C, Kandzia R, Marillonnet S (2008) A one pot, one step, precision cloning method with high throughput capability. *PLoS One* 3 (11). doi:[10.1371/journal.pone.0003647](https://doi.org/10.1371/journal.pone.0003647)
15. Shao Z, Zhao H, Zhao HDNA (2009) Assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res* 37(2):1–10. doi:[10.1093/nar/gkn991](https://doi.org/10.1093/nar/gkn991).
16. Werner S, Engler C, Weber E, Gruetzner R, Marillonnet SA (2012) Modular cloning system for standardized assembly of multigene constructs. *Bioeng Bugs* 3(1):38–43. doi:[10.4161/bbug.3.1.18223](https://doi.org/10.4161/bbug.3.1.18223).
17. Sarrion-Perdigones A, Vazquez-Vilar M, Palací J et al (2013) GoldenBraid 2.0: a comprehensive DNA assembly framework for plant synthetic biology. *Plant Physiol* 162 (3):1618–1631. doi:[10.1104/pp.113.217661](https://doi.org/10.1104/pp.113.217661)
18. Anderson JC, Dueber JE, Leguia M et al (2010) BglBricks: a flexible standard for biological part assembly. *J Biol Eng* 4(1):1. doi:[10.1186/1754-1611-4-1](https://doi.org/10.1186/1754-1611-4-1)
19. De Paoli HC, Tuskan GA, Yang X (2016) An innovative platform for quick and flexible joining of assorted DNA fragments. *Sci Rep* 6:19278. doi:[10.1038/srep19278](https://doi.org/10.1038/srep19278)
20. Casini A, Storch M, Baldwin GS, Ellis T (2015) Bricks and blueprints: methods and standards for DNA assembly. *Nat Rev Mol Cell Biol* 9:1–9. doi:[10.1038/nrm4014](https://doi.org/10.1038/nrm4014)
21. Hutchison CA, Chuang R-YR-Y, Noskov VN et al (2016) Design and synthesis of a minimal bacterial genome. *Science* 351(6280): aad6253. doi:[10.1126/science.aad6253](https://doi.org/10.1126/science.aad6253)
22. Ng DTW, Sarkar CA (2014) NP-Sticky: a web server for optimizing DNA ligation with non-palindromic sticky ends. *J Mol Biol* 426(8):1861–1869. doi:[10.1016/j.jmb.2014.02.003](https://doi.org/10.1016/j.jmb.2014.02.003)

An Adaptive Laboratory Evolution Method to Accelerate Autotrophic Metabolism

Tian Zhang and Pier-Luc Tremblay

Abstract

Adaptive laboratory evolution (ALE) is an approach enabling the development of novel characteristics in microbial strains via the application of a constant selection pressure. This method is also an efficient tool to acquire insights on molecular mechanisms responsible for specific phenotypes. ALE experiments have mainly been conducted with heterotrophic microbes to study, for instance, cell metabolism with different multicarbon substrates, tolerance to solvents, pH variation, and high temperature. Here, we describe employing an ALE method to generate *Sporomusa ovata* strains growing faster autotrophically and reducing CO₂ into acetate more efficiently. Strains developed via this ALE method were also used to gain knowledge on the autotrophic metabolism of *S. ovata* as well as other acetogenic bacteria.

Key words Adaptive laboratory evolution, Autotroph, Acetogen, CO₂ fixation, *Sporomusa ovata*, Methanol, Microbial electrosynthesis

1 Introduction

Adaptive laboratory evolution (ALE) consists in applying a pressure on a microbial population to promote the acquisition of mutations that will improve cell fitness [1]. Most often, ALE experiments are a simple series of culture transfers maintained under a constant stress (Fig. 1) [2]. Part of the culture in the exponential growth phase is transferred to a new tube containing fresh growth medium. Subsequently, the microbial population is continuously transferred under the same condition until the desired characteristics have been developed. This approach enables the enrichment of variants growing faster under the investigated stress. ALE is a powerful technique for the development of beneficial characteristics in microbial strains catalyzing industrial bioprocesses [2]. For instance, ALE has been used to improve production rate [3–5], tolerance to solvents such as biofuels [6–8], tolerance to inhibitors such as acetate found in complex substrates, and tolerance to thermal stress [9]. ALE is also

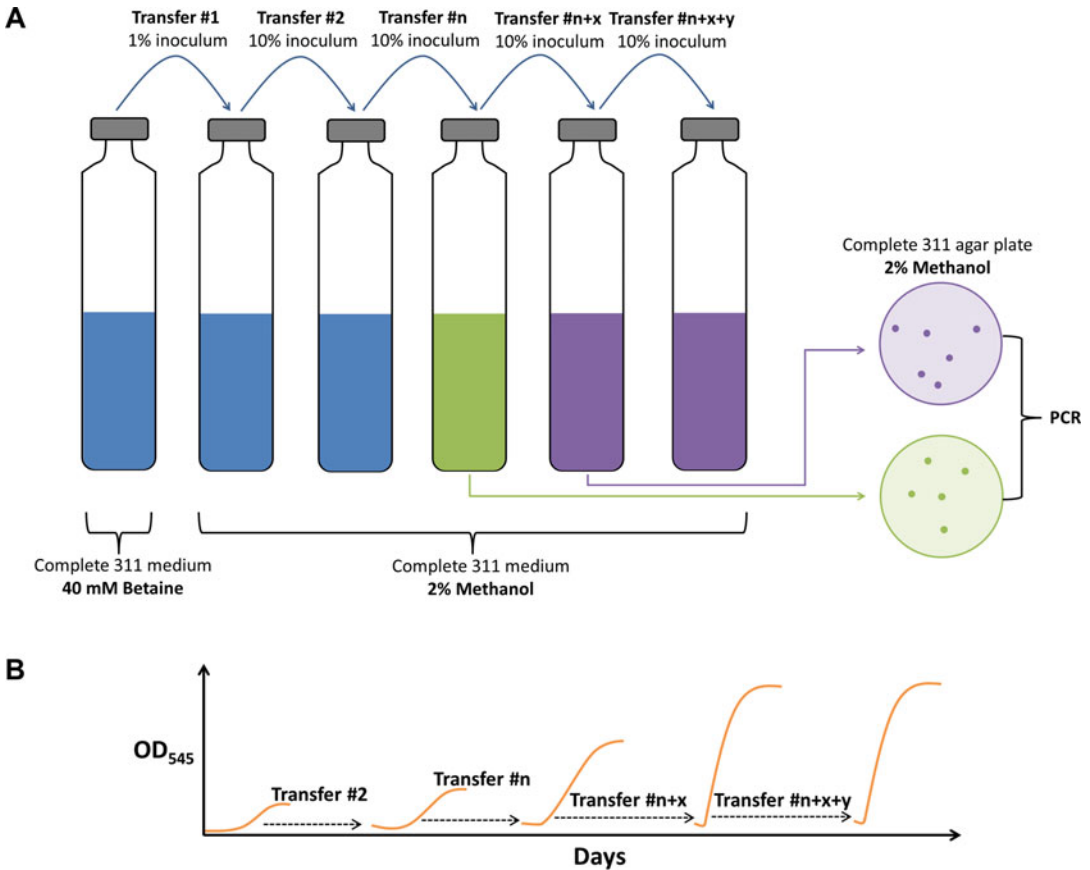


Fig. 1 ALE experiment for the generation of *S. ovata* strains growing faster autotrophically. **(a)** Scheme of sequential transfer in anaerobic culture tubes. The initial *S. ovata* culture is grown heterotrophically with 40 mM betaine before being transferred to complete 311 medium with 2% methanol as the sole substrate. Autotrophic transfers are done with inocula from cultures that are at the beginning of the exponential growth phase. Changes in tube color indicate changes in culture fitness. *Blue* corresponds to initial fitness, *green* corresponds to improved fitness, and *purple* corresponds to optimal fitness. The transfer sequence is stopped when the growth rate does not improve anymore (e.g., transfer # $n + x + y$, n , x and $y > 1$). To isolate clones, transfers of interest (e.g., transfer # n and transfer # $n + x$) are streaked on complete 311 agar plates containing 2% methanol. The identity of isolated clones is verified by PCR. The validated clones will then be subjected to multiple experiments including growth characterization with different substrates, whole-genome sequencing, and RNA sequencing. **(b)** Scheme of growth curve during ALE corresponding to transfer #2, transfer # n , transfer # $n + x$, and transfer # $n + x + y$

an efficient strategy to study molecular mechanisms participating in the development of advantageous phenotypes, stress tolerance, and faster growth rate. Indeed, strains developed by ALE can be characterized with a vast array of molecular tools such as whole-genome sequencing, transcriptomics, and proteomics [2, 10–12]. This combination of techniques generates useful information that can lead to a better understanding of microbial physiology and metabolism.

Recently, ALE has been used to develop strains of *Sporomusa ovata* growing faster under autotrophic conditions and reducing CO₂ more efficiently [11]. In this study, *S. ovata* was adapted to accelerate its growth rate with methanol as the substrate. *S. ovata* is a methylotroph as well as an acetogen that reduces CO₂ into acetate via the Wood-Ljungdahl pathway [13, 14]. During anaerobic methanol oxidation by acetogens, external CO₂ will be used as an electron acceptor and reduced to acetate [15]. In the study by Tremblay et al., the two purposes of ALE were: (1) to develop *S. ovata* strains more efficient at reducing CO₂ for potential applications and (2) to investigate the molecular mechanisms responsible for faster autotrophic growth as well as for higher tolerance to methanol. Beside anaerobic methanol oxidation and gas fermentation (H₂:CO₂), *S. ovata* is also frequently used as a catalyst for microbial electrosynthesis (MES), which is a bioprocess where electrons derived from an electrode are employed by a microbe to reduce CO₂ into multicarbon molecules [16, 17]. Thus, *S. ovata* has the potential to become a microbial catalyst of choice for several promising biotechnologies aiming at converting C1 compounds like methanol and the greenhouse gas CO₂ into multicarbon chemicals of interest.

Here, we will describe in detail the ALE approach that we have employed to generate *S. ovata* variants growing faster under autotrophic conditions. This includes the medium preparation protocol, the anaerobic cultivation techniques that we have applied for both autotrophic and heterotrophic growth, the clone isolation and PCR methods employed to verify the identity of the adapted strains, and the extensive growth characterization strategy with different substrates used to test the adapted strains.

2 Materials

Growth medium and other solutions were all prepared with ultra-pure deionized water. The ultrapure water had a conductivity of 0.055 μS/cm at 25 °C.

2.1 Anaerobic Cultivation

1. Compressed gas cylinders containing pure N₂, N₂:CO₂ (80:20), or H₂:CO₂ (80:20).
2. Balch-type tubes (18 × 150 mm) and serum bottles (100 mL) for anaerobic culture with 20 mm butyl rubber stoppers and aluminum seals.
3. Hand crimper to close anaerobic cultivation tubes and bottles.
4. Coy anaerobic chamber with an oxygen and hydrogen analyzer, heated fan boxes, palladium catalysts, as well as desiccant plates.

**2.2 *Sporomusa ovata*
DSM-2662**

1. *S. ovata* type strain DSM-2662 can be obtained from the Deutsche Sammlung Mikroorganismen und Zellkulturen (DSMZ) (*see Note 1*).
2. The same type strain (ATCC 35899) can also be obtained from the American Type Culture Collection (ATCC).

**2.3 Base 311
Medium and Other
Reagents for *S. ovata*
Cultivation**

1. For base 311 growth medium (DSMZ), add the following reagents to ultrapure water: 0.5 g NH₄Cl, 0.5 g MgSO₄·7H₂O, 0.25 g CaCl₂·2H₂O, 2.25 g NaCl, 0.002 g FeSO₄·7H₂O, 10 mL vitamins solution, 1 mL trace elements solution SL-10, and 1 mL selenite-tungstate solution (*see Note 2*).
2. For vitamin solution, add the following reagents to 800 mL of ultrapure water: 0.002 g biotin, 0.005 g panthothenic acid, 0.0001 g vitamin B-12, 0.005 g *p*-aminobenzoic acid, 0.005 g thioctic acid, 0.005 g nicotinic acid, 0.005 g thiamine, 0.005 g riboflavin, 0.01 g pyridoxine·HCl, and 0.002 g folic acid. Complete volume to 1 L with ultrapure water and store at 4 °C.
3. For trace elements solution SL-10, add the following reagents to 790 mL of ultrapure water: 10 mL FeCl₂·4H₂O (0.5 g) dissolved in HCl (25%), 0.07 g ZnCl₂, 0.1 g MnCl₂·4H₂O, 0.006 g H₃BO₃, 0.19 g CoCl₂·6H₂O, 0.002 g CuCl₂·2H₂O, 0.024 g NiCl₂·6H₂O, 0.036 g Na₂MoO₄·2H₂O. Complete volume to 1 L with ultrapure water and store at 4 °C.
4. For selenite-tungstate solution, add the following reagents to 800 mL of ultrapure water: 0.5 g NaOH, 0.003 g Na₂SeO₃·5H₂O, and 0.004 g Na₂WO₄·2H₂O. Complete volume to 1 L with ultrapure water and store at 4 °C.
5. For the 100× potassium phosphate solution, add the following reagents and complete to 100 mL with ultrapure water: 3.48 g K₂HPO₄ and 2.27 g KH₂PO₄. Prepare the solution anaerobically by bubbling with 100% N₂ gas and store at room temperature after sterilization by autoclaving.
6. For 25× NaHCO₃ solution, dissolve 5 g of NaHCO₃ into 100 mL of ultrapure water. Prepare the solution anaerobically by bubbling with N₂:CO₂ (80:20) gas and store at room temperature after sterilization by autoclaving.
7. For 100× cysteine solution, dissolve 1.76 g cysteine·HCl·H₂O into 80 mL of ultrapure water. Adjust pH to 7.0 and complete with ultrapure water to a final volume of 100 mL. Prepare the solution anaerobically by bubbling with 100% N₂ and store at room temperature after sterilization by autoclaving.

8. For the 50× yeast extract solution, dissolve 5.0 g yeast extract into 100 mL of ultrapure water. Prepare the solution anaerobically by bubbling with 100% N₂ and store at room temperature after sterilization by autoclaving.
9. For the 20× betaine solution, dissolve 10.8 g betaine·H₂O into ultrapure water and adjust pH to 7.0. Complete to 100 mL with ultrapure water. Prepare the solution anaerobically by bubbling with 100% N₂ and store at room temperature after sterilization by autoclaving.
10. Prepare an anaerobic 100% methanol solution by bubbling with 100% N₂.
11. Noble agar and petri dishes.
12. Dimethyl sulfoxide (DMSO).

2.4 Colony PCR and DNA Gel

1. DreamTaq DNA polymerase (5 U/μL) (ThermoFisher Scientific).
2. 10× DreamTaq buffer with 20 mM MgCl₂ (ThermoFisher Scientific).
3. dNTP mix, 2 mM of dATP, dCTP, dGTP, and dTTP (ThermoFisher Scientific).
4. Forward primer (10 μM solution) for the detection of *S. ovata* clone: SOpyrEIN-1; 5'-GTTGGTATGCTGGTTGACCGCA GTGG-3'.
5. Reverse primer (10 μM solution) for the detection of *S. ovata* clone: SopyrEdn; 5'-CCATCCAGGCAAATAGTAACGAG CCA-3'.
6. 0.2 mL thin-walled PCR tubes.
7. S1000 thermal cycler (Bio-Rad).
8. 1% Agarose gel stained with SYBR Safe DNA Gel Stain (ThermoFisher Scientific).
9. 50× TAE buffer. For its preparation, add 242 g Tris-base, 57.1 mL 100% acetic acid, and 100 mL 0.5 M EDTA solution pH 8.0 to 600 mL of ultrapure water, and then complete to 1 L with ultrapure water.
10. GeneRuler 1 kb DNA ladder (ThermoFisher Scientific).
11. 6× DNA Gel loading dye (ThermoFisher Scientific).
12. Electrophoresis gel box.
13. Electrophoresis power supply.
14. Safe Imager 2.0 Blue Light Transilluminator (ThermoFisher Scientific).

3 Methods

3.1 311 Medium Preparation

1. Prepare base 311 medium and dispense into Balch-type tubes or serum bottles. Bubble base 311 medium with N₂:CO₂ (80:20) and sterilize by autoclaving.
2. After autoclaving, under anaerobic and sterile conditions add pre-prepared 10 mL 100× potassium phosphate solution, 40 mL 25× NaHCO₃ solution, and 10 mL 100× cysteine solution to a volume of base 311 medium sufficient to obtain 1 L of complete 311 medium after the addition of the required substrate (*see Note 3*).
3. Optionally, under anaerobic and sterile conditions add pre-prepared 20 mL 50× yeast extract solution for every liter of complete 311 medium.
4. For heterotrophic growth, under anaerobic and sterile conditions add pre-prepared 50 mL 20× betaine solution for every liter of complete 311 medium.
5. For autotrophic growth with methanol as the substrate, under anaerobic and sterile conditions add either pre-prepared 5 mL (0.5%) or 20 mL (2%) 100% methanol for every liter of complete 311 medium.
6. For autotrophic growth with H₂:CO₂ (80:20), replace the N₂:CO₂ atmosphere in anaerobic tubes or serum bottle containing complete 311 medium by bubbling and pressurizing to 1.7 atm.

3.2 Reactivation of *S. ovata* DSM-2662 Strain from Freezer Via Heterotrophic Cultivation

1. *S. ovata* strains are stored at −80 °C in 10% DMSO.
2. Use 1 mL of *S. ovata* DSM-2662 from −80 °C storage freezer to inoculate an anaerobic tube containing 10 mL complete 311 medium amended with 40 mM betaine and 0.1% yeast extract (*see Note 4*).
3. Incubate the heterotrophic *S. ovata* culture at 30 °C until reaching an optical density at 545 nm (OD₅₄₅) of ca. 1.0.

3.3 ALE of *S. ovata* with Methanol as the Substrate (Fig. 1)

1. Use 0.1 mL of *S. ovata* DSM-2662 cultivated heterotrophically with betaine to inoculate anaerobic tubes containing 10 mL complete 311 medium amended with 2% methanol (*see Note 5*).
2. Incubate the 2% methanol *S. ovata* cultures at 30 °C until reaching an OD₅₄₅ between 0.1 and 0.2. This is considered as transfer #1 of the ALE experiment (*see Notes 6–8*).
3. Use 10% of transfer #1 culture to inoculate new anaerobic cultivation tubes containing fresh 311 medium amended with 2% methanol. This is transfer #2.

4. Repeat **step 2**.
5. During the ALE experiment, *S. ovata* culture are transferred repeatedly as described in **steps 2–4** until growth rate with 2% methanol reached a plateau and cannot be improved anymore [11].
6. A fraction of each transfer cultures should be stored in 10% DMSO at -80°C for potential subsequent studies.

3.4 Isolation of Clones from ALE Cultures (See Note 9)

1. Prepare anaerobically agar plates made with complete 311 medium amended with 2% methanol.
2. Add 15% noble agar to base 311 medium and bubble with N_2 : CO_2 gas before sterilization by autoclaving.
3. Cool down base 311 agar medium at 55°C and amend with pre-prepared potassium phosphate solution, NaHCO_3 solution, cysteine solution, and 2% methanol. Solution addition should be carried out under anaerobic and sterile solution.
4. In an anaerobic chamber set at 30°C and filled with a N_2 : CO_2 : H_2 (76.5–78:20:2–3.5) atmosphere, pour complete 311 medium with 2% methanol into Petri dishes. Wait until the next day to use these agar plates to avoid condensation. For *S. ovata* optimal growth in the anaerobic chamber, CO_2 concentration should be 20% and H_2 concentration should be 2–3.5% (see **Note 10**).
5. Isolate clones from transfer cultures of interest by streaking on agar plates containing complete 311 medium with 2% methanol.
6. Incubate the plates in second containment in the anaerobic chamber until the apparition of isolated colonies that can easily be picked up for colony PCR and for liquid medium inoculation.

3.5 Colony PCR to Validate the Identity of *S. ovata* Variants Generated by ALE

1. Prepare on ice 50 μL PCR reactions in 0.2 mL thin-walled PCR tubes.
2. For each 50 μL PCR reaction, add 5 μL 10 \times DreamTaq buffer with 20 mM MgCl_2 , 5 μL 2 mM dNTPs, 5 μL 10 μM forward primer, 5 μL 10 μM reverse primer, and 0.5 μL DreamTaq DNA polymerase (5 U/ μL).
3. In the anaerobic chamber, touch a *S. ovata* colony grown on a complete 311 agar plate amended with 2% methanol with the tip of a toothpick. Dip it into the PCR reaction.
4. Place the prepared PCR reactions in a S1000 thermal cycler and start the following PCR cycle: 3 min at 95°C , 30 s at 94°C , 30 s at 58°C , 1 min and 10 s at 72°C , 10 min at 72°C . Repeat **steps 2–4** for 35 times.

5. After the PCR cycle, add 10 μL 6 \times DNA loading dye to each 50 μL PCR reaction tube.
6. Load 5–10 μL of each PCR reaction to a 1% SYBR-Safe stained agarose gel immersed in 1 \times TAE buffer in an electrophoresis gel box. On the same gel, load one well with 5 μL of GeneRuler 1 kb DNA ladder.
7. Run the 1% agarose gel at 100 V until the dye line reaches $\frac{3}{4}$ of the total gel.
8. Put the gel on a Safe Imager 2.0 Blue Light Transilluminator to detect DNA band. If the colony is formed by *S. ovata* cells and is not a contamination, a single PCR band of ca. 1.1 kb should be detected.
9. For each ALE transfer of interest, touch at least three positive *S. ovata* colonies with the tip of a needle attached to a syringe and inoculate them into anaerobic tubes containing 10 mL complete 311 medium amended with 2% methanol.
10. Incubate anaerobic cultivation tubes at 30 °C.
11. After the cultures grew, prepare DMSO freezer stock for future experiments.
12. These cultures are the isolated clones from different transfers of interest that will be used for subsequent experiments including growth characterization, whole-genome resequencing, and RNA sequencing (*see* **Notes 11** and **12**).

**3.6 Growth
Characterization
During and After the
ALE Experiment (See
Note 13)**

1. During the ALE, growth of the transferred cultures or of clones isolated from transferred cultures can be characterized with different substrates including betaine, $\text{H}_2:\text{CO}_2$ atmosphere, and different concentrations of methanol. The autotrophic metabolism of *S. ovata* cultures can also be evaluated for microbial electrosynthesis where the required electrons are derived from an electrode and CO_2 is the only carbon source (*see* **Note 14**).
2. For growth on different substrates, inoculate a given volume of the *S. ovata* culture to fresh 311 medium in triplicate to get an initial OD₅₄₅ of ca. 0.02.
3. For growth characterization with betaine as the substrate, inoculate anaerobic tubes containing 10 mL complete 311 medium amended with 40 mM betaine but without yeast extract (*see* **Note 15**).
4. For growth characterization with methanol as the substrate, inoculate anaerobic tubes containing 10 mL complete 311 medium amended with different concentrations of methanol (e.g., 0.5% and 2%).

5. For growth characterization with $\text{H}_2:\text{CO}_2$ atmosphere, inoculate anaerobic tubes containing 10 mL complete 311 medium with a $\text{H}_2:\text{CO}_2$ atmosphere (1.7 atm).

4 Notes

1. *S. ovata* is an acetogenic bacterium that grows heterotrophically or autotrophically with substrates including betaine, fructose, $\text{H}_2:\text{CO}_2$, and methanol [13]. *S. ovata* autotrophic metabolism can also be driven with electrons coming from the cathode electrode of a MES reactor [16, 18, 19]. One interesting feature of *S. ovata* is that it can grow in a defined medium, which does not require the addition of complex ingredients such as yeast extract or peptone solutions. This characteristic facilitates ALE studies aiming at accelerating growth rate with a specific substrate, since complex ingredients could contain alternative substrates that would make results interpretation difficult.
2. The final volume of the 311 base medium will depend on the type of substrate that will be used for *S. ovata* growth. The final volume of complete 311 medium should be 1 L.
3. The potassium phosphate solution and the NaHCO_3 solution are added to base 311 medium after autoclaving to avoid the formation of a precipitate.
4. Inoculation and cultivation of *S. ovata* are handled under anaerobic and sterile conditions. Syringes used to inoculate bacterial cultures or to inject solutions to base 311 medium to complete 311 medium must be flushed beforehand with $\text{N}_2:\text{CO}_2$. Cysteine is added to base 311 medium as a reducing agent to remove residual oxygen.
5. ALE experiments with *S. ovata* or with other acetogens aiming at improving growth rate under autotrophic conditions can also be conducted via sequential transfer with inoculum from the exponential growth phase with other substrates than methanol such as a $\text{H}_2:\text{CO}_2$ atmosphere. In the study by Tremblay et al., 2015, 2% methanol was chosen as the substrate for ALE because it is also a toxic solvent for bacteria [11, 20, 21]. Bacteria have been shown to reduce their exposure to toxic compounds by metabolizing them [22]. Thus, the toxicity of methanol creates an additional pressure on the *S. ovata* population for the selection of mutant cells breaking down methanol more efficiently and possibly growing autotrophically faster. The other reason why 2% methanol was used as the substrate for ALE by Tremblay et al. 2015 is that one of the goals of the study was to gain knowledge on the molecular mechanism conferring solvent tolerance in acetogenic bacteria.

6. The inoculum volume of transfer #1 from a heterotrophic *S. ovata* culture using betaine as a substrate to complete 311 medium amended with 2% methanol is low (1%) compared with the following transfers (10%). The purpose is to reduce as much as possible the quantity of unused betaine that will be carried on in the autotrophic cultivation medium of transfer #1.
7. The ALE experiment is performed in triplicate. In transfer #1, three anaerobic cultivation tubes filled with complete 311 medium amended with 2% methanol are inoculated from a heterotrophic *S. ovata* culture. Subsequently, transfer #1-1, #1-2, and #1-3 will be the starting points for three series of independent sequential transfers.
8. During ALE with 2% methanol, *S. ovata* are always transferred when the cultures reached an OD₅₄₅ of ca. 0.1–0.2. These ODs correspond to the entire exponential growth phase for unadapted *S. ovata* and to the beginning of the exponential growth phase for *S. ovata* cells already adapted to 2% methanol [11].
9. ALE cultures contain a mixed population of genetic variants that have acquired different sets of mutation during the sequential transfer process. Some of these mutations increase the fitness of microbial cells while others are silent. Clonal colonies must be isolated for subsequent experiments such as whole-genome resequencing and RNA sequencing aiming at understanding molecular mechanisms responsible for fitness increase. Another purpose of colony isolation is to verify that ALE cultures after many transfers only comprise *S. ovata* cells without contamination.
10. CO₂ concentration is maintained at 20% in the anaerobic chamber as well as in cultivation tubes or serum bottles because 311 medium pH of ca. 7.0 is maintained via a bicarbonate buffer system. H₂ concentration is maintained at 2–3.5% in the anaerobic chamber to react with the palladium catalyst ensuring the removal of residual O₂. Higher H₂ concentrations than 4% generate risks of explosion.
11. Because it is optional, method for whole-genome sequencing will not be described in detail here. Briefly, genomic DNA of clones isolated from ALE transfers of interest is extracted with Easy-DNA gDNA purification kit (ThermoFisher Scientific) following the manufacturer's instructions. Paired-end libraries of genomic DNA used for sequencing are prepared with TruSeq Nano DNA LT Sample Preparation Kit (Illumina). Whole-genome sequencing is done on a MiSeq (Illumina) platform with a MiSeq Reagent kit v2 (Illumina) with a paired-end protocol and read lengths of 151 nucleotides. Trimmomatic

is used to trim sequencing reads and *breseq* is employed for variants calling. The genome sequence of the strain *Sporomusa ovata* DSM 2662 (NCBI accession ASXP00000000.1), which is the starting strain of the ALE experiment described here, is used as reference [14]. All the investigated samples should have coverage of at least 30×.

12. Because it is optional, the method for RNA sequencing will not be described in detail here. Briefly, *S. ovata* strains investigated for differential gene expression (e.g., wild type vs. ALE-evolved strains) are grown in triplicate under identical growth conditions. When bacterial cultures reached the exponential phase, cells are centrifuged and resuspended in Max Bacterial Enhancement Reagent (ThermoFisher Scientific) to improve RNA extraction efficiency. Total RNA is then extracted with TRIzol Max Bacterial RNA Isolation Kit (ThermoFisher Scientific) and purified further with RNeasy Mini Kit (Qiagen) with on-column DNase treatment. To enrich the messenger RNAs fraction and significantly reduce the quantity of present ribosomal RNAs, total RNA is treated with a Ribo-Zero rRNA Removal kit (Illumina). Subsequently, sequencing libraries are prepared with a TruSeq RNA Sample Preparation kit (Illumina). Sequencing is done on a MiSeq (Illumina) platform with a MiSeq Reagent kit v2 (Illumina). RNA sequencing data are treated with Rockhopper 2.03 [23] for reads mapping, normalization, and transcript abundance quantification. The genome of *S. ovata* DSM 2662 is used as the reference for reads mapping [14].
13. Two of the most frequently used approaches to characterize the growth and the metabolism of acetogens are to monitor cell density evolution by measuring optical density with a spectrophotometer and to measure the production of acetate as well as the consumption of substrate via High Performance Liquid Chromatography. Acetate is the main product synthesized by acetogens during both heterotrophic and autotrophic growth [15].
14. Because it is optional, the method for MES will not be described in detail here. Briefly, *S. ovata* cultures to be tested for MES are grown in triplicate in the cathode chamber of three-electrode, H-type bioelectrochemical reactors as described previously [16]. In the H-type reactor, the cathode chamber is separated from the anode chamber by a Nafion 115 proton-exchange membrane, which enables the transit of protons generated by water splitting at the anode to the cathode compartment. The Nafion 115 membrane also limits O₂ transfer to the cathode, ensuring that strict anaerobes such as *S. ovata* are not exposed to aerobic conditions. Electrons flow from the anode via an electric circuit and are delivered to the

cathode. A potentiostat is employed to control the cathode potential at a specific potential through a reference electrode (e.g., -690 mV versus Standard Hydrogen Electrode or -900 mV versus Ag/AgCl). The potentiostat is also used to collect electrochemical data. The anode and cathode electrodes are usually made of carbonaceous material such as graphite. Both the anode and cathode electrodes are suspended in 250 mL of complete 311 medium without electron donor, cysteine, and yeast extract. During the MES experiment, the two chambers of the reactor are constantly stirred with magnetic stirrers and bubbled with $N_2:CO_2$ (80:20) to maintain pH, anaerobic conditions, dissolved CO_2 concentration, and uniformity. At the beginning of the MES experiment, the cathode chamber will be inoculated with a *S. ovata* culture that usually has been pre-grown autotrophically with $H_2:CO_2$. Current draw, OD_{545} , the production of acetate and of other compounds coming from the reduction of CO_2 will be monitored throughout the MES experiment.

15. For heterotrophic growth characterization, no yeast extract is added to complete 311 medium amended with betaine to ensure that the only available substrate for growth is betaine and that the medium is completely defined.

References

1. Elena SF, Lenski RE (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat Rev Genet* 4:457–469. doi:[10.1038/nrg1088](https://doi.org/10.1038/nrg1088)
2. Dragosits M, Mattanovich D (2013) Adaptive laboratory evolution – principles and applications for biotechnology. *Microb Cell Factories* 12:64. doi:[10.1186/1475-2859-12-64](https://doi.org/10.1186/1475-2859-12-64)
3. Patzschke A, Steiger MG, Holz C et al (2015) Enhanced glutathione production by evolutionary engineering of *Saccharomyces cerevisiae* strains. *Biotechnol J* 10:1719–1726. doi:[10.1002/biot.201400809](https://doi.org/10.1002/biot.201400809)
4. Zambanini T, Sarikaya E, Kleineberg W et al (2016) Efficient malic acid production from glycerol with *Ustilago trichophora* TZ1. *Biotechnol Biofuels* 9:67. doi:[10.1186/s13068-016-0483-4](https://doi.org/10.1186/s13068-016-0483-4)
5. Hu B, Yang Y-M, Beck DAC et al (2016) Comprehensive molecular characterization of *Methylobacterium extorquens* AM1 adapted for l-butanol tolerance. *Biotechnol Biofuels* 9:84. doi:[10.1186/s13068-016-0497-y](https://doi.org/10.1186/s13068-016-0497-y)
6. Horinouchi T, Tamaoka K, Furusawa C et al (2010) Transcriptome analysis of parallel-evolved *Escherichia coli* strains under ethanol stress. *BMC Genomics* 11:579. doi:[10.1186/1471-2164-11-579](https://doi.org/10.1186/1471-2164-11-579)
7. Atsumi S, Wu T-Y, Machado IMP et al (2010) Evolution, genomic analysis, and reconstruction of isobutanol tolerance in *Escherichia coli*. *Mol Syst Biol* 6:449. doi:[10.1038/msb.2010.98](https://doi.org/10.1038/msb.2010.98)
8. Reyes LH, Almario MP, Winkler J et al (2012) Visualizing evolution in real time to determine the molecular mechanisms of n-butanol tolerance in *Escherichia coli*. *Metab Eng* 14:579–590. doi:[10.1016/j.ymben.2012.05.002](https://doi.org/10.1016/j.ymben.2012.05.002)
9. Oide S, Gunji W, Moteki Y et al (2015) Adaptive laboratory evolution conferred cross-tolerance to thermal and solvent stress to *Corynebacterium glutamicum*. *Appl Environ Microbiol*. doi:[10.1128/AEM.03973-14](https://doi.org/10.1128/AEM.03973-14)
10. Tremblay P-L, Summers ZM, Glaven RH et al (2011) A *c*-type cytochrome and a transcriptional regulator responsible for enhanced extracellular electron transfer in *Geobacter sulfurreducens* revealed by adaptive evolution. *Environ Microbiol* 13:13–23. doi:[10.1111/j.1462-2920.2010.02302.x](https://doi.org/10.1111/j.1462-2920.2010.02302.x)
11. Tremblay P-L, Höglund D, Koza A et al (2015) Adaptation of the autotrophic acetogen

- Sporomusa ovata* to methanol accelerates the conversion of CO₂ to organic products. Sci Rep 5:16168. doi:[10.1038/srep16168](https://doi.org/10.1038/srep16168)
12. LaCroix RA, Sandberg TE, O'Brien EJ et al (2015) Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of *Escherichia coli* K-12 MG1655 on glucose minimal medium. Appl Environ Microbiol 81:17–30. doi:[10.1128/AEM.02246-14](https://doi.org/10.1128/AEM.02246-14)
 13. Möller B, Oßmer R, Howard BH et al (1984) *Sporomusa*, a new genus of gram-negative anaerobic bacteria including *Sporomusa sphaeroides* spec. nov. and *Sporomusa ovata* spec. nov. Arch Microbiol 139:388–396. doi:[10.1007/BF00408385](https://doi.org/10.1007/BF00408385)
 14. Poehlein A, Gottschalk G, Daniel R (2013) First insights into the genome of the Gram-negative, endospore-forming organism *Sporomusa ovata* strain H1 DSM 2662. Genome Announc 1:e00734–e00713. doi:[10.1128/genomeA.00734-13](https://doi.org/10.1128/genomeA.00734-13)
 15. Drake HL, Gößner AS, Daniel SL (2008) Old acetogens, new light. Ann N Y Acad Sci 1125:100–128. doi:[10.1196/annals.1419.016](https://doi.org/10.1196/annals.1419.016)
 16. Nevin KP, Woodard TL, Franks AE et al (2010) Microbial electrosynthesis: feeding microbes electricity to convert carbon dioxide and water to multicarbon extracellular organic compounds. mBio 1:e00103–e00110. doi:[10.1128/mBio.00103-10](https://doi.org/10.1128/mBio.00103-10)
 17. Tremblay P-L, Zhang T (2015) Electrifying microbes for the production of chemicals. Front Microbiol 6:201. doi:[10.3389/fmicb.2015.00201](https://doi.org/10.3389/fmicb.2015.00201)
 18. Chen L, Tremblay P-L, Mohanty S et al (2016) Electrosynthesis of acetate from CO₂ by a highly structured biofilm assembled with reduced graphene oxide–tetraethylene pentamine. J Mater Chem A 4:8395–8401. doi:[10.1039/C6TA02036D](https://doi.org/10.1039/C6TA02036D)
 19. Aryal N, Halder A, Tremblay P-L et al (2016) Enhanced microbial electrosynthesis with three-dimensional graphene functionalized cathodes fabricated via solvothermal synthesis. Electrochim Acta 217:117–122. doi:[10.1016/j.electacta.2016.09.063](https://doi.org/10.1016/j.electacta.2016.09.063)
 20. Schrader J, Schilling M, Holtmann D et al (2009) Methanol-based industrial biotechnology: current status and future perspectives of methylotrophic bacteria. Trends Biotechnol 27:107–115. doi:[10.1016/j.tibtech.2008.10.009](https://doi.org/10.1016/j.tibtech.2008.10.009)
 21. Patterson JA, Ricke SC (2015) Effect of ethanol and methanol on growth of ruminal bacteria *Selenomonas ruminantium* and *Butyrivibrio fibrisolvens*. J Environ Sci Health B 50:62–67. doi:[10.1080/03601234.2015.965639](https://doi.org/10.1080/03601234.2015.965639)
 22. Nicolaou SA, Gaida SM, Papoutsakis ET (2010) A comparative view of metabolite and substrate stress and tolerance in microbial bioprocessing: from biofuels and chemicals, to biocatalysis and bioremediation. Metab Eng 12:307–331. doi:[10.1016/j.ymben.2010.03.004](https://doi.org/10.1016/j.ymben.2010.03.004)
 23. McClure R, Balasubramanian D, Sun Y et al (2013) Computational analysis of bacterial RNA-Seq data. Nucleic Acids Res 41:e140. doi:[10.1093/nar/gkt444](https://doi.org/10.1093/nar/gkt444)

Chapter 11

CRISPR-Cas9 Toolkit for Actinomycete Genome Editing

Yaojun Tong, Helene Lunde Robertsen, Kai Blin, Tilmann Weber,
and Sang Yup Lee

Abstract

Bacteria of the order Actinomycetales are one of the most important sources of bioactive natural products, which are the source of many drugs. However, many of them still lack efficient genome editing methods, some strains even cannot be manipulated at all. This restricts systematic metabolic engineering approaches for boosting known and discovering novel natural products. In order to facilitate the genome editing for actinomycetes, we developed a CRISPR-Cas9 toolkit with high efficiency for actinomycetes genome editing. This basic toolkit includes a software for spacer (sgRNA) identification, a system for in-frame gene/gene cluster knockout, a system for gene loss-of-function study, a system for generating a random size deletion library, and a system for gene knockdown. For the latter, a uracil-specific excision reagent (USER) cloning technology was adapted to simplify the CRISPR vector construction process. The application of this toolkit was successfully demonstrated by perturbation of genomes of *Streptomyces coelicolor* A3(2) and *Streptomyces collinus* Tü 365. The CRISPR-Cas9 toolkit and related protocol described here can be widely used for metabolic engineering of actinomycetes.

Key words CRISPR-Cas9, CRISPRi, Uracil-specific excision reagent (USER) cloning, Synthetic biology, Actinomycetes, Genome editing, Double-strand break (DSB), Homology directed repair (HDR), Non-homologous end joining (NHEJ)

1 Introduction

1.1 Actinomycetes

Actinomycetes are Gram-positive bacteria with high GC content genomes, belonging to the order of Actinomycetales. They are well known for their ability to produce medically and industrially relevant secondary metabolites (natural products) [1–3], including, but not limited to antibiotics, herbicides, chemotherapeutics, and immunosuppressants, such as vancomycin, bialaphos, doxorubicin, and rapamycin, respectively. However, after being studied over half-century, it becomes more and more challenging to find novel secondary metabolites with meaningful properties by traditional

The authors have filed a patent (EP15160126.7) on the actinomycete CRISPR toolkit.

methods. However, modern genome mining techniques [4–7] have revealed that those bacteria still possess a huge unexploited potential to produce secondary metabolites with novel structures [8]. Unfortunately, in comparison with model organisms like *E. coli* and *S. cerevisiae*, there are only few genetic manipulation tools available for actinomycetes. In addition, the high GC content (sometimes >72%) impedes genetic manipulation even if actinomycete DNA is manipulated in other hosts like *E. coli*. With the help of the recently developed CRISPR-Cas9 technology, we now have more tools to address and overcome these challenges for efficient genetic manipulation of actinomycetes.

1.2 CRISPR-Cas9

The modules of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/CRISPR-associated (Cas) proteins are present in most archaea and many bacteria as adaptive immune systems for defense against foreign DNA [9–11] or RNA [12]. Based on the number of Cas proteins involved, CRISPR-Cas systems can be divided into “class 1” and “class 2.” Class 1 systems have multiple Cas9 proteins, while class 2 systems only need one single Cas protein, for instance, type II CRISPR system, also known as CRISPR-Cas9 system [13]. The currently well-studied and widely used CRISPR-Cas9 system is originally from *Streptococcus pyogenes*. The target sequences of the Cas9 endonuclease are defined in the CRISPR loci containing short repeats separated by “spacer” sequences that exactly match the sequences of the targeted foreign genetic element. Introducing double-strand breaks (DSBs) into these DNAs offers adaptive immunity against foreign genetic elements [9, 14–17]. In the native CRISPR-Cas9 system, the spacer sequence of the CRISPR array transcribes to a CRISPR RNA (crRNA). Subsequently, an associated trans-activating CRISPR RNA (tracrRNA) hybridizes with the crRNA, forming an RNA duplex, which is cleaved and further processed by endogenous RNase III and possibly other, yet unknown nucleases [18]. The crRNA-tracrRNA duplex, which was later artificially designed as a chimera named “single guide RNA” (sgRNA) [17], interacts with Cas9 to form a complex, then scans the foreign genetic elements for the presence of trinucleotide protospacer adjacent motifs (PAMs). When this complex finds a PAM that has a 5' sequence (normally around 20 nt) complementary to the spacer sequence in the crRNA-tracrRNA duplex, it binds to this position and then triggers the conformational change of Cas9 to activate the HNH and RuvC endonuclease domains [19, 20], which causes DNA double-strand break (DSB). The DNA DSB of a chromosome is lethal, and cells can only survive if the lesion is repaired. The two major routes for DNA repair are (1) non-homologous end joining (NHEJ), in which no editing template is needed, and (2) homology-directed repair (HDR), in which an editing template for homologous recombination is needed [21, 22] (Fig. 1).

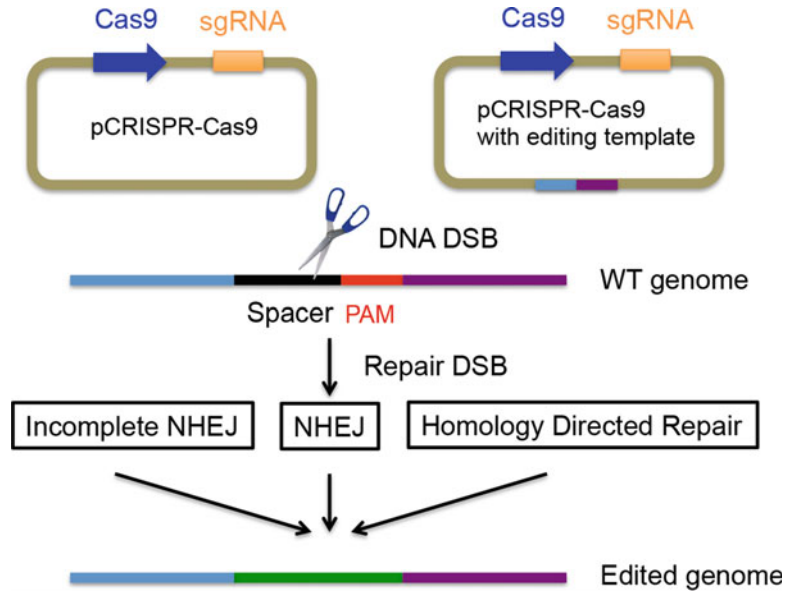


Fig. 1 CRISPR-Cas9 schematics. CRISPR-Cas9 genome editing working model

Using the feature of introducing DNA DSBs, which then get repaired by cellular mechanisms such as NHEJ or HDR, allows the use of CRISPR-Cas9 for genome editing in many different organisms, from *E. coli* to human cells [23–31]. NHEJ is a complicated process that involves several proteins. It has been best described in eukaryotic cells to repair the DNA DSB in an error-prone manner [32, 33]; however, it is also found in prokaryotes [34, 35]. Interestingly, in *S. coelicolor* we found that the NHEJ in contrast to other bacteria, e.g., *Mycobacterium tuberculosis* [34], is missing a DNA ligase function. In this strain, the activity of this enzyme can be partially restored by other yet unknown enzymes, albeit with a lower efficiency. In *S. coelicolor* and other streptomycetes lacking the ligase gene, the native “incomplete” NHEJ repair pathway usually leads to larger deletions around the site of the DSB. This phenomenon can be exploited to trigger deletions between the nearest two essential genes, thus leading to “random size deletion libraries” [36, 37].

A study of *S. pyogenes* Cas9 nuclease domain revealed that mutating the HNH and RuvC domains (D10A and H840A) resulted in a catalytically dead Cas9 (dCas9) variant that does not have endonuclease activity, but could still form a complex with sgRNA and efficiently bind to the target DNA [17]. This effect can be used to sequence-specifically interfere with transcription and thus control gene expression. In analogy to eukaryotic RNA interference (RNAi), this system was named as CRISPRi [38].

1.3 USER Cloning

Construction of CRISPR-Cas9 vectors using ligation-based approach is still relatively time consuming, and is difficult to be implemented in high-throughput and automation settings.

Nowadays, PCR-based cloning is a commonly used method for de novo gene assembly in metabolic engineering [39]. USER friendly cloning is one of those modern cloning methods, and it represents an alternative to conventional ligation-based cloning in that it allows for simultaneous scarless assembly of multiple PCR products into USER-compatible vectors (Fig. 2). This allows for easy and versatile vector construction [40]. We introduced USER friendly cloning to facilitate CRISPR-Cas9 vectors construction, as well as to meet the demands for further high-throughput and automated genome editing purposes.

USER assembly relies on the generation of complementary overhangs in the PCR products and destination vector and can be divided into three distinct steps. First, genes of interest (GOIs) are PCR amplified with primers containing between 7 and 12 nucleotides overhangs flanked by uracil bases (dU) [41]. The directional assembly of the PCR fragments is facilitated through the overhangs, which are designed either manually or using an online tool such as AMUSER 1.0 (at <http://www.cbs.dtu.dk/services/AMUSER/>) [42]. In addition, a proofreading DNA polymerase such as PfuX7 (Norholm, *see* [43]) or the commercially available Phusion U Hot Start DNA Polymerase (Thermo Fisher Scientific, Waltham, US) is required for the recognition of the uracil bases and incorporation of adenosine residues on the complementary strand; second, the destination vector is linearized and with that complementary overhangs generated using a combination of a restriction and a nicking enzyme. The enzymes required for linearization and generation of single-stranded overhangs depend on the USER cassette in the destination vector. Examples of USER cassettes include the PacI/Nt.BbvCI, AsiSI/Nb.BsmI, and AsiSI/Nb.BtsI cassettes [44]; in the third step, PCR fragments are assembled in the linearized vector by means of the USER™ kit (New England Biolabs) that contains a mixture of the *E. coli* uracil DNA glycosylase and DNA glycosylase-lyase endonuclease VIII, both of which recognize and remove uracil bases. Following uracil excision, the reaction is kept at the melting temperature of the single-stranded overhangs for several minutes to facilitate the assembly of the PCR fragments in the destination vector [45].

During the past 2 years, independent laboratories have established modular and efficient genetic manipulation tools for streptomycetes based on CRISPR-Cas9. These tools significantly facilitated the processes of gene/gene cluster deletion, point mutagenesis, gene replacement, as well as repression of gene transcription in *Streptomyces* [36, 46–48]. In this chapter, we describe protocols using the toolkit developed in our lab [36, 37] and a workflow combining of in silico primer design for sgRNA construction, USER-based cloning, and CRISPRi.

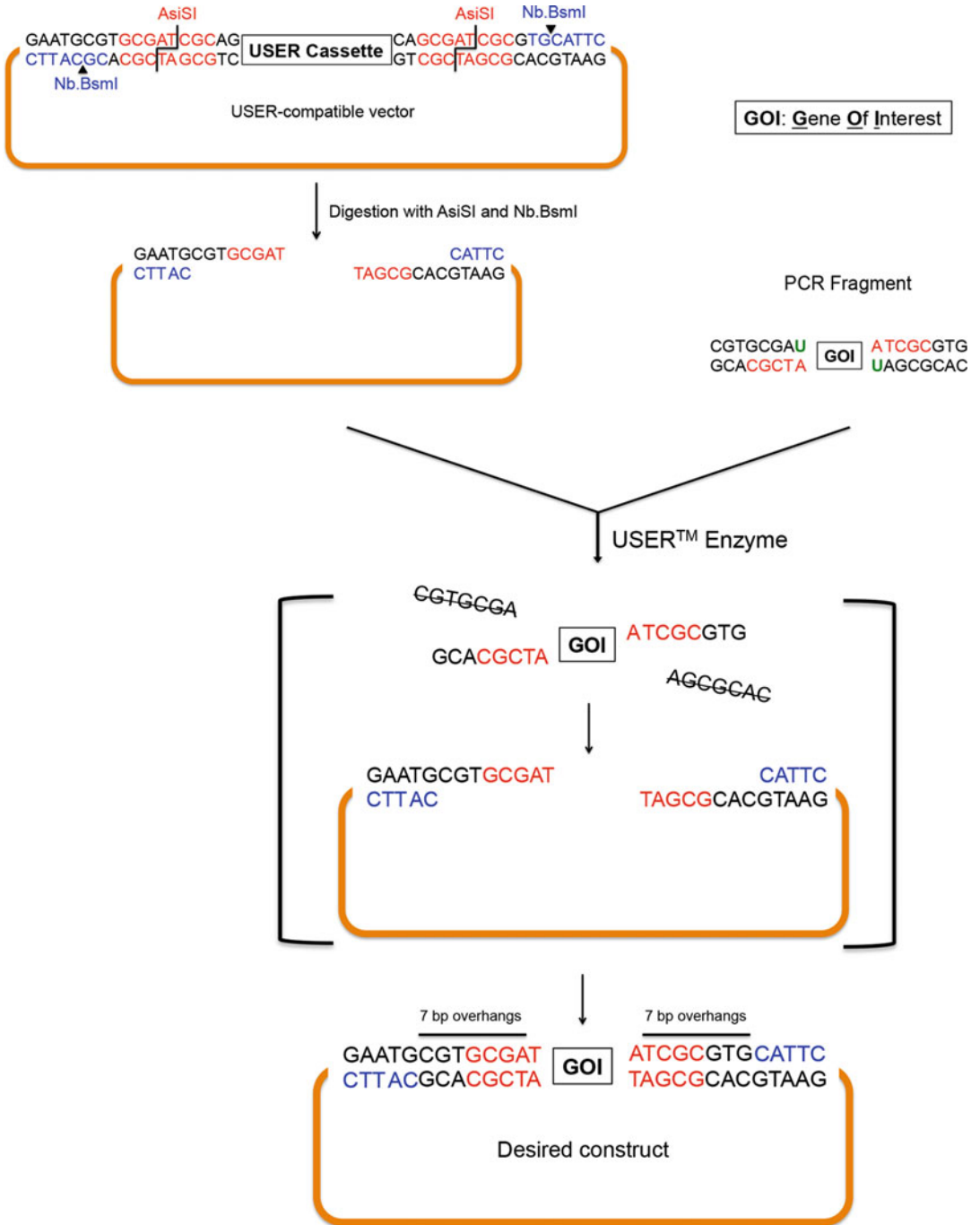


Fig. 2 USER cloning schematics. Schematic overview of the steps involved in USER assembly of PCR-generated fragments in a linearized USER-compatible destination vector

2 Materials

Milli-Q water (18.2 M Ω cm at 25 °C) is used for preparation of all media and solutions. All kits and reagents are used according to the manufacturer's instructions, unless the modifications are indicated. Diligently follow all waste disposal regulations when disposing waste materials.

2.1 Strains

One Shot[®] ccdB Survival[™] 2 T1^R chemically competent cells (Thermo Fisher Scientific) are used for the construction of the USER-compatible vectors. Chemically competent *E. coli* cells, e.g., NEB5- α (New England Biolabs), and One Shot[®] Mach1[™] T1 Phage-Resistant (Thermo Fisher Scientific) are used for routine cloning. Non-methylating *E. coli* ET12567/pUZ8002 [49] is used for conjugation. *Streptomyces coelicolor* A3(2), and *Streptomyces collinus* Tü365 are used as example strains in these protocols.

2.2 Plasmids

Plasmids pGM1190 [50], pGM1190-Cas9, pCRISPR-Cas9, pCRISPR-dCas9, pCRISPR-Cas9-ScaligD [36], and pCRISPR-USER-(d)Cas9 are used in the following protocols. All oligonucleotides and gBlocks are purchased from Integrated DNA Technologies (IDT).

2.3 Media

All components for media preparation are purchased from Sigma-Aldrich, unless indicated otherwise.

SOC medium (20 g/L Tryptone, 5 g/L Yeast extract, 4.8 g/L MgSO₄, 3.603 g/L Dextrose, 0.5 g/L NaCl, 0.186 g/L KCl), LB medium (10 g/L Tryptone, 5 g/L Yeast extract, 5 g/L NaCl, 20 g/L Agar is added for solidification), ISP2 medium (Yeast extract 4 g/L, Malt extract 10 g/L, Dextrose 4 g/L, 20 g/L Agar is added for solidification), and Soya Flour Mannitol agar (MS, or SFM, or Cullum agar) (20 g/L Mannitol, 20 g/L Soya flour with low fat (W. Schoenenberger GmbH & Co.), 20 g/L Agar) supplemented with 10 mM MgCl₂. Appropriate antibiotics are added to the media when needed. Their working concentrations are: apramycin, 50 μ g/mL; nalidixic acid, 50 μ g/mL; thiostrepton, 1 μ g/mL; kanamycin, 25 μ g/mL; and chloramphenicol, 25 μ g/mL.

2.4 Reagents and Kits

Phusion Hot Start II DNA Polymerase (2 U/ μ L), DreamTaq Green PCR Master Mix (2 \times), PCR Master Mix (2 \times), PfuX7 DNA polymerase [43], Phusion U Hot Start DNA Polymerase (2 U/ μ L), T4 DNA Ligase (1 U/ μ L), GeneJET Plasmid Miniprep Kit, CloneJET PCR Cloning Kit, GeneJET PCR Purification Kit, and all restriction enzymes are purchased from Thermo Fisher Scientific; USER[™] Enzyme, and Gibson Assembly[®] Cloning Kit are purchased from New England Biolabs. Blood & Cell Culture

DNA Kit is from Qiagen. NucleoSpin[®] Gel Clean-up kit is from Macherey-Nagel.

2.5 Equipment

NanoDrop 2000 (Thermo Scientific) is used to measure DNA concentrations, and Concentrator plus (Eppendorf) is used for concentrating DNA solutions.

3 Methods

All the procedures are carried out at room temperature unless otherwise specified, all DNAs are eluted by nuclease-free water (pH 8). The annealing temperature (T_a) is calculated with the Thermo Fisher Scientific T_m calculator: <http://www.thermofisher.com/dk/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/tm-calculator.html>.

The plasmids within our CRISPR-Cas9 toolkit are based on one single temperature sensitive vector pGM1190, which is a derivative of the replicon pSG5 [50]. All necessary elements are integrated into one single construct, where Cas9/dCas9 is under control of the thiostrepton inducible *tipA* promoter, while the sgRNA is driven by a constitutive *ermE** promoter. The editing template, NHEJ missing component(s), and other element(s) can be inserted via the singular StuI site of the vector [36] (*see Note 1*).

3.1 Identification of Suitable 20 nt Spacers with CRISPy-Web

For successful CRISPR/Cas9 experiments it is essential to define good 20 nt spacer sequences within the desired target region of the genome. One prerequisite for these 20 nt spacers is that they match the 20 nt upstream of a PAM close to the desired target. In addition, it is important to avoid off-target effects: If the same or a very similar 20 nt spacer sequence is found close to a PAM elsewhere on the genome, Cas9 will introduce a DSB there as well. To reduce the probability of unwanted side effects even further, spacers that match many other spacers with a mismatch of one or two bases should be avoided.

For this reason, it is recommended to use computational tools to design the guide RNAs. While many programs exist for designing sgRNAs for model organisms, only few tools can be used with user-supplied genomes [51]. The CRISPy-web tool [52] assists researchers in this task by identifying appropriate 20 nt spacer regions for sgRNAs in any user-supplied microbial genome sequence. CRISPy-web is available at <http://crispy.secondarymetabolites.org/>.

1. To run CRISPy-web for a genome region of interest, a GenBank-formatted file can be uploaded by clicking the “Browse” button and selecting the appropriate file on the

start page of CRISPy-web. Alternatively, CRISPy-web supports directly using the results from the antiSMASH secondary metabolite biosynthetic gene cluster mining platform [4, 5], by simply selecting “Get sequences from antiSMASH” and providing the antiSMASH job id that is included in the antiSMASH result email. The search is started by clicking the “Start” button.

2. Once the sequence has been uploaded, the user has to select a target region to scan for suitable spacers. The region selection page gives a short summary of the uploaded genome and a search field below the summary can be used to specify the target region.

Regions can be selected by entering a range of nucleotide coordinates (like 12,345–67,890), a gene name, or a locus tag from the annotated genome.

The search field will autocomplete for gene and cluster name as well as locus tags. If the data were directly transferred from antiSMASH and secondary metabolite biosynthetic gene cluster has been identified, these are also shown in a summary table below the search field.

A help screen for the syntax is displayed by clicking on “Usage hints.”

3. Once a target region is selected, a click on the “Find targets” button starts the actual scan for spacers. Depending on the size of the selected region and the overall genome size, this process usually takes between few—for small—to around 15 min for large genomes.
4. Once the scan has completed, the user will be presented with a page showing an overview of the scanned region. In this overview, genes are displayed as gray arrows and potential 20 nt spacers are indicated as red boxes. It is possible to zoom in to a specific gene by clicking the gene arrow and selecting “Show results for this gene only” from the pop-up.
5. Potential spacers are displayed sorted by the number of identical off-target hits, the number of off-target hits allowing for one or two mismatches, and the location on the genome. When hovering over the table, the currently active spacer is highlighted in the visualization, and vice versa.
6. Clicking on a table row adds the corresponding spacer to the download basket, a second click deselects the spacer again. The basket icon on the upper right of the screen displays the number of selected spacers.
7. Clicking the download basket icon takes the user to the download page. Here, a summary of the selected spacers is shown, and it is possible to download the selection as a comma-separated table that can be opened from a spreadsheet application or text editor.

3.2 Generation of a Random Size Deletion Library in Actinomycetes Using pCRISPR-Cas9 with Native NHEJ (See Note 2)

1. Digest the pCRISPR-Cas9 with Fast Digest NcoI and SnaBI or Eco105I restriction enzyme (Thermo Fisher Scientific, Waltham, US, Waltham, US). To prepare a stock of vector, a 100 μ L reaction is used. Mix 5 μ g (up to 10 μ g) plasmid, 10 μ L 10 \times Fast Digest Buffer, 5 μ L of each of the Fast Digest restriction enzymes, and nuclease-free Milli-Q water to 100 μ L. Incubate at 37 $^{\circ}$ C for 60 min. The digested plasmid is purified by GeneJET PCR Purification Kit (*see Note 3*). Then use NanoDrop 2000 to measure the concentration, and then Concentrator plus for concentrating the DNA solution when needed. The NcoI and SnaBI double-digested pCRISPR-Cas9 backbone solution can be stored in small aliquots at -20° C for up to 6 months for multiple usages.
2. Identify spacers for functional sgRNA cassettes using CRISPy-web tool as described in Subheading 3.1; for each gene of interest, pick two spacers with minimal off-target effects.
3. Design primers for functional sgRNA cassette amplification, the forward primer can be designed as sgRNA-F:
5'-CATGCCATGGN₂₀GTTTTAGAGCTAGAAATAGC-3'
(N₂₀ represents the 20 nt spacer sequence);
the reverse primer stays the same as sgRNA-R:
5'-ACGCCTACGTAAAAAAGCACCGACTCGGTGCC-3'.
The restriction enzyme sites are underlined.
4. PCR is used to amplify the functional sgRNA cassette from pCRISPR-Cas9. 50 μ L PCR reaction is used. Mix 20 ng (up to 100 ng) plasmid DNA, 10 μ L 5 \times HF Buffer, 1 μ L 10 mM dNTP mix, 0.5 μ M of designed primers, 1.5 μ L DMSO, 1 U Phusion Hot Start II DNA Polymerase, and nuclease-free Milli-Q water to 50 μ L on ice, flip the PCR tubes by fingers, spin down the mixture. The PCR conditions are 98 $^{\circ}$ C for 30 s; 35 cycles of 98 $^{\circ}$ C for 10 s; T_a (up to 72 $^{\circ}$ C, calculated by Thermo Fisher Scientific T_m calculator from both primers) for 30 s; 72 $^{\circ}$ C for 10 s (1 kb/15–30 s); and finally 72 $^{\circ}$ C for 10 min, afterward keep at 4 $^{\circ}$ C.
5. Analyze the PCR products using 2% agarose gel on 1 \times TAE running buffer, the positive PCR product is purified by GeneJET PCR Purification Kit. Then use NanoDrop 2000 to measure the concentration, and Concentrator plus for concentrating the DNA solution when needed.
6. The purified PCR products are double digested by Fast Digest NcoI and SnaBI restriction enzymes, with the same condition of the plasmid double digestion.
7. Mix 100 ng of the double-digested pCRISPR-Cas9 backbone from **step 1**, fivefold of double-digested functional sgRNA cassette PCR product (from **step 6**), 1 μ L 10 \times T4 Buffer,

and 1 U T4 DNA ligase in a total of 10 μL reaction volume, then incubate at 25 °C for 60 min.

8. Transform 50 μL of One Shot[®] Mach1™ T1 Phage-Resistant competent *E. coli* cells with 10 μL the ligation mixture, heat shock at 42 °C for 70 s, recover the cells in 300 μL SOC medium at 37 °C, 200 rpm for 1 h. Plate 200 μL of the recovered cells on selective LB agar plates with 50 $\mu\text{g}/\text{mL}$ apramycin, and incubate at 37 °C overnight (around 16 h).
9. On the next day, pick 3–5 colonies into 0.5 mL selective LB liquid medium with 50 $\mu\text{g}/\text{mL}$ apramycin in 1.5 mL Eppendorf tubes, incubate at 37 °C, 200 rpm for 4 h. 1 μL of each culture is used as a template for colony PCR validation of the ligation in **step 7**.
10. The colony PCR is carried out in a 20 μL reaction using Taq-based DNA polymerase. Mix 1 μL of the culture from **step 8** (20 ng of non-digested pCRISPR-Cas9 as a negative control), 10 μL of the PCR Master Mix (2 \times), 0.5 μM of the primers (sgRNA check-F: 5'-AATTGTACGCGGTCGATCTT-3' and sgRNA check-R: 5'-TACGTAAAAAAGCACCGAC-3'), and nuclease-free Milli-Q water to 20 μL on ice, flip the PCR tubes by fingers, spin down the mixture. Colony PCR conditions are 94 °C for 4 min; 35 cycles of 94 °C for 30 s; 50 °C for 30 s; 72 °C for 20 s (1 kb/1 min); and finally 72 °C for 10 min, then keep at 4 °C.
11. Analyze the PCR products using 4% agarose gel (*see Note 4*) on 1 \times TAE running buffer. Randomly pick two positive clones of each construct for 10 mL overnight culture using selective LB liquid medium with 50 $\mu\text{g}/\text{mL}$ apramycin, at 37 °C.
12. On the next day, isolate the plasmids from the 10 mL culture of **step 10**, and confirm the results by Sanger sequencing using primer sgRNA check-F.
13. Transform ET12567/pUZ8002 competent *E. coli* cell with 100 ng of the validated plasmid using the same protocol as in **step 8**. Plate 100 μL of the recovered cells on selective LB agar plates with 50 $\mu\text{g}/\text{mL}$ apramycin, 50 $\mu\text{g}/\text{mL}$ kanamycin, and 25 $\mu\text{g}/\text{mL}$ chloramphenicol, and incubate at 37 °C for around 24 h.
14. Randomly pick one clone (known as the donor strain for conjugation) from the selective LB plate of **step 12**, inoculate it into 10 mL of the same selective LB liquid medium, incubate at 37 °C overnight (around 24 h).
15. Wash the above culture twice with 10 mL LB liquid medium without antibiotics supplementation and then suspend the cell pellet with 1 mL (1/10 volume of the culture) LB liquid medium.

16. Mix 100 μL ET12567/pUZ8002 culture from **step 15** with 50 μL *S. coelicolor* A3(2) spores (*see Note 5*), and plate the mixture onto Cullum agar plates, inoculate the plates at 30 °C overnight (around 16 h).
17. On the next day, overlay the conjugation plates with 1 mL of sterilized Milli-Q water containing 1 mg nalidixic acid and 1 mg apramycin.
18. Incubate the plates at 30 °C for 3–5 days to let the exconjugants grow.
19. Pick customized number of exconjugants (the library size) and re-streak them onto ISP2 plates with 1 $\mu\text{g}/\text{mL}$ thiostrepton, 50 $\mu\text{g}/\text{mL}$ apramycin, and 50 $\mu\text{g}/\text{mL}$ nalidixic acid for 5–7 days (*see Note 6*).
20. Inoculate the **step 19** strain into 20 mL non-antibiotic ISP2 liquid medium and incubate at 30 °C, 180 rpm for 3–5 days.
21. Isolate genomic DNA of the strains from **step 20** using Blood & Cell Culture DNA Kit.
22. The isolated genomic DNA can be used to analyze the random size deletion library.

**3.3 Highly Efficient
Gene Loss-of-Function
Studies in
Actinomycetes
Using pCRISPR-Cas9-
ScaligD (See Note 7)**

Steps of 1–18 are identical to Subheading 3.2, except the backbone plasmid is pCRISPR-Cas9-ScaligD instead of pCRISPR-Cas9. The protocol diverges at **step 19**, when exconjugants can be seen from the conjugation plates.

19. Pick 10–20 exconjugants and re-streak them onto ISP2 plates with 1 $\mu\text{g}/\text{mL}$ thiostrepton, 50 $\mu\text{g}/\text{mL}$ apramycin, and 50 $\mu\text{g}/\text{mL}$ nalidixic acid for 5–7 days.
20. Scratch some mycelia of the clones from **step 19** using a sterile toothpick into 10 μL pure DMSO (Sigma-Aldrich, St. Louis, US) in PCR tubes. Shake tubes vigorously for 10 min at 100 °C in a shaking heating block, vortex vigorously for another 1 min at room temperature, spin down the pellet at top speed for 10 s. Then 2 μL of the supernatant is used as a PCR template in a 50 μL reaction in **step 22**.
21. Design primers for amplifying an approximately 500 bp fragment, around the expected DSB site of mutation.
22. Mix 2 μL supernatant from **step 20**, 10 μL 5 \times GC Buffer, 1 μL 10 mM dNTP mix, 0.5 μM of designed primers from **step 21**, 1 U Phusion Hot Start II DNA Polymerase and nuclease-free Milli-Q water to 50 μL on ice, flip the PCR tubes by fingers, spin down the mixture. The PCR conditions are 98 °C for 30 s; 35 cycles of 98 °C for 10 s; T_a (up to 72 °C, is calculated by Thermo Fisher Scientific T_m calculator from both the primers) for 30 s; 72 °C for 10 s (1 kb/15–30 s); and finally 72 °C for 10 min, 4 °C forever.

23. Subclone the PCR products of **step 22** into pJET1.2/blunt vector from CloneJET PCR Cloning Kit, use pJET1.2 Forward Sequencing Primer from the kit for Sanger sequencing.
24. The desired mutations from **step 23** can be used for gene loss-of-function study.

3.4 Generation of In-Frame Gene/Gene Cluster Deletions or Replacements in Actinomycetes Using pCRISPR-Cas9 with Homologous Recombination Templates

As the homologous recombination is widely used for gene deletion and replacement in many organisms including *Streptomyces* [53]. We provide an editing template within the same plasmid, pCRISPR-Cas9 for HDR of the DSB caused by Cas9, to achieve scar-less genome editing.

Almost all the steps are the same as Subheading 3.2 (steps 1–11 are exactly the same). The main different steps of this protocol are homologous recombination templates design and cloning.

12. Digest the pCRISPR-Cas9 that carries the designed spacer (the plasmid from **step 11** of Subheading 3.3) with Fast Digest StuI or Eco147I restriction enzyme. A 100 μ L reaction volume is used. Mix 5 μ g (up to 10 μ g) plasmid, 10 μ L 10 \times Fast Digest Buffer, and 5 μ L of Fast Digest StuI restriction enzyme. Then incubate at 37 $^{\circ}$ C for 30 min. The linearized plasmid is purified by GeneJET PCR Purification Kit. Then use NanoDrop 2000 to measure the concentration and Concentrator plus for concentrating the DNA solution when needed.
13. Design primers for amplifying \sim 1 kb of both 5' and 3' fragments of the target gene (gene cluster), with 20 nt overhang at the end of both the fragments for later Gibson Assembly.
14. PCR amplify around 1 kb homologous recombination templates from genomic DNA of the WT strain. The 50 μ L PCR reaction is used. Mix 150 ng (up to 500 ng) genomic DNA, 10 μ L 5 \times GC Buffer, 1 μ L 10 mM dNTP mix, 0.5 μ M of designed primers, 1.5 μ L DMSO, 1 U Phusion Hot Start II DNA Polymerase, and nuclease-free Milli-Q water to 50 μ L on ice, flip the PCR tubes by fingers, spin down the mixture. The PCR conditions are 98 $^{\circ}$ C for 30 s; 35 cycles of 98 $^{\circ}$ C for 10 s; T_a (up to 72 $^{\circ}$ C, is calculated by Thermo Fisher Scientific T_m calculator from both the primers) for 30 s; 72 $^{\circ}$ C for 30 s (1 kb/15–30 s); and finally 72 $^{\circ}$ C for 10 min, then keep at 4 $^{\circ}$ C.
15. A 3-fragment Gibson assembly of 10 μ L reaction volume is used to assemble the two homologous recombination templates into StuI site of pCRISPR-Cas9 with designed spacer. Mix 100 ng linearized plasmid (backbone), three- to fivefold of each of the two \sim 1 kb homologous recombination templates, 5 μ L 2 \times Gibson Master Mix, and nuclease-free Milli-Q water to 10 μ L in a PCR tube on ice. Flip the tube, spin down the mixture, incubate at 50 $^{\circ}$ C in a PCR block for 60 min.

16. Transform 50 μL of One Shot[®] Mach1[™] T1 Phage-Resistant competent *E. coli* cells with 10 μL of each assembly reaction. Apply heat shock at 42 °C for 70 s, recover the cells in 300 μL SOC medium at 37 °C for 1 h. Plate 200 μL of the recovered cells on selective LB agar plates with 50 $\mu\text{g}/\text{mL}$ apramycin, and incubate at 37 °C overnight (around 16 h).
17. On the next day, pick 3–5 colonies into 0.5 mL selective LB liquid medium with 50 $\mu\text{g}/\text{mL}$ apramycin in 1.5 mL Eppendorf tubes, incubate at 37 °C, 200 rpm for 4 h for colony PCR using the designed primers which can cross both homologous recombination templates to validate the assembly in **step 15**.
18. The colony PCR is carried out in a 20 μL reaction, the same protocol as **step 9** of Subheading 3.3. Check the PCR products using 1% agarose gel on 1 \times TAE running buffer. Randomly pick two positive clones of each construct for 10 mL overnight culture. The following steps are the same as **steps 11–18**.
19. Randomly pick 3–5 exconjugants and restreak them onto ISP2 plates with 1 $\mu\text{g}/\text{mL}$ thiostrepton, 50 $\mu\text{g}/\text{mL}$ apramycin, and 50 $\mu\text{g}/\text{mL}$ nalidixic acid for 5–7 days.
20. Inoculate the **step 19** strains into the 20 mL non-antibiotic ISP2 liquid medium and incubate at 40 °C, 180 rpm for 7 days to eliminate the CRISPR plasmid (*see Note 8*).
21. A proper (can be 1000- to 10,000-fold, depends on the culture density) diluted fraction of **step 20** cultures is plated on non-antibiotic ISP2 plates to isolate single colonies.
22. Randomly pick 10–20 colonies of each strain from **step 21** and replica plate on ISP2 agar with and without 50 $\mu\text{g}/\text{mL}$ apramycin. The clones with restored apramycin sensitivity have successfully eliminated the temperature-sensitive pCRISPR-Cas9 plasmid with the homologous recombination templates.
23. Scratch some mycelia of the plasmid-free clones from **step 22** using a sterile toothpick into 10 μL pure DMSO in PCR tubes. Let the tubes shake vigorously for 10 min at 100 °C in a shaking heating block, vortex vigorously for 1 min at room temperature, spin down the pellet at top speed for 10 s, 2 μL of the supernatant is used as the PCR template in a 50 μL reaction in **step 24** (*see Note 9*).
24. Mix 2 μL supernatant from **step 23**, 10 μL 5 \times GC Buffer, 1 μL 10 mM dNTP mix, 0.5 μM of designed primers from **step 17**, 1 U Phusion Hot Start II DNA Polymerase, and nuclease-free Milli-Q water to 50 μL on ice, flip the PCR tubes with fingers, spin down the mixture. The PCR conditions are 98 °C for 30 s; 35 cycles of 98 °C for 10 s; T_a (annealing temperature, up to 72 °C, was calculated by Thermo Fisher Scientific T_m calculator from the primers) for 30 s; 72 °C for X seconds (depends on

the amplicon size) (1 kb/15–30 s); and finally 72 °C for 10 min, keep on 4 °C.

25. Subclone the PCR products of **step 24** into pJET1.2/blunt vector from CloneJET PCR Cloning Kit, and use pJET1.2 Forward Sequencing Primer from the kit for Sanger sequencing.
26. The desired in-frame deletions can be identified from **step 25** (*see Note 10*).

3.5 pCRISPR-dCas9 for Gene Knockdown in *Actinomyces*

For bacterial CRISPRi application, there are two locations of the DNA that can be targeted by dCas9:sgRNA complex to suppress the transcription; the gene coding region, and the region upstream of the start codon, which often includes the promoter. However, caused by a yet unknown mechanism, for sgRNAs targeting the coding region, only those sgRNAs that bind to the non-template DNA strand have the suppression effect, while no such strand bias is observed when the sgRNAs are targeting on the 5'-UTR [36, 38]. A simple illustration is shown in Fig. 3a, while a detailed example is in Fig. 3b [36].

The steps are the same as **steps 1–19** of Subheading 3.2. Except the spacer from the coding region of the sgRNAs needs to target on the non-template DNA strand.

20. The strains from **step 19** can be used for gene knock-down validation, either by measuring the final product or the mRNA level.

3.6 Introduction of USER Cassette (See Notes 11 and 12) into pCRISPR Serials Vector for the Facilitation of the CRISPR Vector Construction

The vector pGM1190-Cas9 is used to exemplify the de novo construction of a USER-compatible vector. In general, the USER vector contains a USER cassette composed of the *ccdB* gene-chloramphenicol resistance marker flanked by a restriction and a nicking enzyme. For the CRISPR-Cas9 system, thiostrepton and apramycin resistance markers are included in the vector for positive selection in *actinomyces* hosts. Two USER-compatible vectors are currently available in our lab: the pCRISPR-USER-Cas9 for gene knockout (in), and the pCRISPR-USER-dCas9 for gene knockdown.

1. The *ccdB*-chloramphenicol resistance marker cassette is amplified from the Gateway[®] Vector of the “Gateway[®] Vector Conversion System with One Shot[®] *ccdB* Survival Cells”-kit. For the PCR, mix 20 μL 5× Phusion[®] HF Buffer, 2 μL 10 mM dNTP mix, 1 μM of each primer (5'-AAAACGCGGCGGAATGCGTGCGATCGCAG-3' and 5'-AAAAGGGCCCGAATGCACGCGATCGCTG-3') (*see Note 13*), 2 U Phusion[®] HF DNA Polymerase, and nuclease-free Milli-Q water to 100 μL on ice, flip the PCR tubes by fingers, spin down the mixture. The PCR conditions are 98 °C for 30 s; 35 cycles of

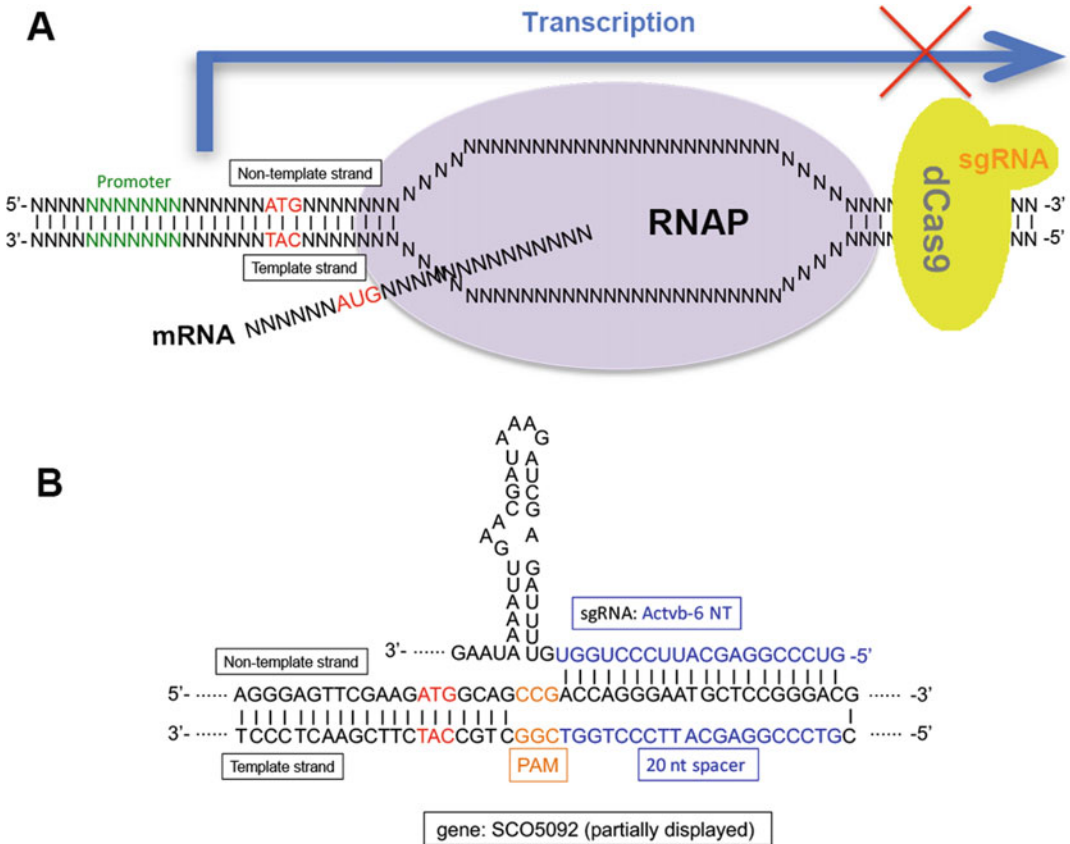


Fig. 3 Prokaryotic CRISPRi working model. (a) An overview of the prokaryotic CRISPRi working model. (b) Specific example of CRISPRi, the sgRNA is targeting on (binding to) the nontemplate DNA strand of the coding region of the SCO5092 gene in *S. coelicolor* [36]

98 °C for 10 s; 71 °C for 30 s; 72 °C for 1 min; and finally 72 °C for 10 min, keep on 4 °C.

2. Analyze the PCR product on 1% agarose gel with 1× TAE running buffer and purify the fragment using a gel cleanup kit.
3. Digest vector pGM1190-Cas9 and the PCR-amplified USER cassette with MreI and BseI20I. To ensure sufficient yields, digest up to 10 µg of the destination vector with 5 U of each enzyme in a total volume of 100 µL. For the PCR-amplified USER cassette, a 50 µL reaction with up to 3 µg DNA and 2 U of each enzyme normally is sufficient. The digestions are carried out at 37 °C. It is recommended to run the digestions overnight.
4. Analyze the digestions on 1% agarose gel with 1× TAE running buffer and purify the fragments with a gel cleanup kit.
5. For the ligation, mix digested vector and USER cassette in the ratio 1:3 (see **Note 14**). Mix 2 µL T4 DNA Ligase Reaction

Buffer, 1 U T4 DNA Ligase, and nuclease-free Milli-Q water in a total volume of 20 μ L. The reaction can be carried out either at 25 °C for 1 h or 16 °C overnight and should be terminated by heating at 65 °C for 10 min.

6. Use 10 μ L of the ligation mix for transformation of 50 μ L One Shot[®] ccdB Survival[™] 2 T1^R chemically competent cells. To account for variations in transformation efficiency, plate both 100 and 250 μ L on pre-warmed selective LB plates with 50 μ g/mL apramycin and 25 μ g/mL chloramphenicol.
7. Verify clones by colony PCR and Sanger sequencing with primers covering the entire USER cassette (*see Note 15*). For the colony PCR, use 2 μ L 10 \times DreamTaq buffer, 0.4 μ L 10 mM dNTP mix, 0.1 μ M of each primer (*see Note 15*), 2 μ L colony (dissolved in 20 μ L nuclease-free Milli-Q water), 0.2 U DreamTaq DNA Polymerase, and nuclease-free Milli-Q water to 20 μ L on ice, flip the PCR tubes by fingers, spin down the mixture. PCR conditions are 95 °C for 3 min; 40 cycles of 95 °C for 30 s; 64 °C for 30 s; 72 °C for 2 min; and finally 72 °C for 5 min, keep on 4 °C. Analyze colony PCRs on 1% agarose with 1 \times TAE running buffer. Successful integration of the USER cassette should result in a PCR product of ~1.9 kb. Colonies with correct size of PCR product are subjected to Sanger sequencing. The final vector is named pCRISPR-USER-Cas9.
8. Linearize between 10 and 15 μ g USER-compatible pCRISPR-USER-Cas9 with 20 U and 10 U Nb.BsmI in a total volume of 50 μ L. Add 5 μ L of the NEB3.1 or CutSmart[®] buffer and run the digestion at 37 °C for 2 h, followed by the inactivation of the enzymes at 65 °C for 1 h.
9. Run the 50 μ L-reaction directly on 1% agarose gel with 1 \times TAE running buffer until a clear separation of the linearized vector and the *ccdB* + chloramphenicol resistance marker cassette (size of 1.7 kb).
10. Purify the vector with a gel cleanup kit (*see Note 16*).
11. GOIs are amplified using USER-compatible primers specific for the AsiSI/Nb.BsmI USER cassette. 7 to 12 nucleotides USER overhangs is sufficient for successful assembly [41]:
 5'-CGTGCGAU-[GOI_1]-3'.
 5'-CACGCGAU-[GOI_1]-3'.

Use 20 ng gBlock DNA, and up to 200 ng genomic DNA as PCR template. For the PCR, mix 10 μ L 5 \times Phusion GC Buffer with 1 μ L 10 mM dNTP mix, 0.1 μ M of each oligonucleotide, 1 U PfuX7 (*see Note 17*) or 1 U Phusion U Hot Start DNA Polymerase, and nuclease-free Milli-Q water to 50 μ L on ice, flip the PCR tubes by fingers, spin down the mixture. The

PCR conditions are 98 °C for 3 min; 40 cycles of 98 °C for 10 s; T_a (up to 72 °C, is calculated by Thermo Fisher Scientific T_m calculator from both the primers) for 50 s; 72 °C for 10 s (1 kb/15–30 s); and finally 72 °C for 10 min, keep on 4 °C.

12. Analyze the PCR products on 1% agarose gel with 1× TAE running buffer and purify the right using a gel cleanup kit (*see Note 18*).
13. Mix the linearized USER-compatible vector with PCR fragments in the ratio 1:3 or 1:6. For PCR fragments <1000 bp use ratio 1:6 and for PCR fragments >1000 bp use ratio 1:3. Mix 1 µL USER™ enzyme mix and 0.5 µL 10× CutSmart® buffer with nuclease-free Milli-Q water in a total volume of 10 µL.
14. The USER excision is carried out at 37 °C for 15 min followed by 15 min at T_m of the USER overhang (if using AMUSER for USER overhang prediction software tools, then the T_m is 26 °C). Keep the mix at 10 °C for another 10 min to allow the DNA fragments be assembled into the vector.
15. For transformation, mix all 10 µL USER reaction with 90 µL NEB5-α competent *E. coli* cells and incubate the mixture on ice for 20 min before performing the heat-shock at 42 °C for 45 s. Incubate the transformation mixture on ice for 2 min followed by the addition of 250 µL pre-warmed (37 °C) SOC medium for recovering at 37 °C, 250 rpm for 1 h. Plate 100 and 250 µL (*see Note 19*) of the recovered cells on pre-warmed (37 °C) selective LB plates with appropriate antibiotics (in our case 50 µg/mL apramycin) and incubate at 37 °C overnight (about 16 h).
16. Perform colony PCR as described in **step 7**.

4 Notes

1. Before using the CRISPR-Cas9 toolkit in the strain of interest, please check that the pGM1190-based plasmid can replicate and that the strain codes for a *tipA* homologue that is required to activate the *tipA*-promoter.
2. This feature can only be used in the host with a defective NHEJ. The strain of interest needs to be checked first if there is a defective NHEJ and which component(s) are missing. For example, *S. coelicolor* A3(2) does not contain a gene encoding the ligase LigD.
3. The NcoI and SnaBI double digestion of pCRISPR-Cas9 generates two fragments, ~11 kb and 90 bp, because 90 bp is too small to be caught by the GeneJET PCR Purification Kit, we normally directly use GeneJET PCR Purification Kit to purify

the ~11 kb backbone, which will give you a higher yield. Of course, gel purification can be used to recycle the ~11 kb backbone.

4. Because the differences of successful ligation and self-ligation or the non-digested pCRISPR-Cas9 plasmid PCR are only 20 bp, which needs a high concentration of agarose gel to distinguish.
5. The amounts of ET strain and spore are highly case dependent, in this case, for *S. coelicolor* A3(2), 10^8 ET strains and 10^7 – 10^8 spore per plate are sufficient to generate enough exconjugants, for other actinomyces strains, the conjugation condition needs to be customized.
6. Because the *tipA* promoter is leaky, we have observed that in some cases no induction with thioestrepton is required to provide sufficient Cas9 in the targeted cells.
7. This system, pCRISPR-Cas9-ScaligD, can only be functional in the host that is lack of DNA ligase component of NHEJ pathway.
8. For those actinomycetes that are high temperature sensitive: To eliminate the CRISPR plasmid, carry out **step 20** under 30 °C at least two times and then plate a proper diluted fraction of the culture on nonselective ISP2 plate to isolate single colonies. Check for apramycin sensitivity of these clones. If no candidate shows apramycin sensitivity, additional rounds of nonselective culture may be applied.
9. If colony PCR did not give you any positive bands, which may indicate that the colony PCR did not work, the genomic DNA needs to be isolated for PCR template.
10. If the two homologous recombination templates are outside the gene cluster, then this system can be used for the deletion of whole gene clusters/genomic regions. Alternatively, they can be designed to generate in frame deletions within single genes or used to introduce additional genes/DNA fragments at the targeted DNA region.
11. Instead, a simpler USER cassette can be designed in which the *ccdB* and chloramphenicol resistance marker cassette is omitted. For a detailed protocol on how to construct this type of cassette, see [54].
12. We use the *ccdB*-chloramphenicol resistance marker cassette as positive control for digestion but other systems might apply. The cassette can be amplified from the Gateway[®] Vector found in the Gateway[®] Vector Conversion System with One Shot[®] *ccdB* Survival Cells.
13. For de novo design of the USER cassette, design oligonucleotides with overhangs for *AsiSI* and *Nb.BsmI* as follows:

5'-AAAA-[MCS_enzyme1]-[Nb.BsmI]-[nt*]-[AsiSI]-3'.

5'-AAAA-[MCS_enzyme2]-[Nb.BsmI]-[nt*]-[AsiSI]-3'.

*The variable nucleotide (nt) ensures directional assembly during cloning.

14. To calculate the molar ratios, use a ligation calculator such as the one found at <http://nebiocalculator.neb.com/#/>.
15. Examples of colony PCR and sequencing primers for the pCRISPR-Cas9/dCas9 constructs:
 - USERseq_F: 5'-CGAGCGTCCGCCGGCG-3'.
 - USERseq_internal_F: 5'-GCTAGTGTTCATAGTCCTGAAAATCATCTG-3'.
 - USERseq_internal_R: 5'-CTGGGTGAGTTTCACCAGTTT TGATTTAAAC-3'.
 - USERseq_R: 5'-GCGTACCGCTTCGGGCCC-3'.

Use USERseq_F and USERseq_R for colony PCRs and USERseq_internal_F and USERseq_internal_R for Sanger sequencing of the USER cassette.

16. To improve titers, elute DNA in smaller volumes and repeat the last elution step.
17. The PfuX7 DNA polymerase is prepared in-house by expressing the his-tagged protein followed by Ni-NTA chromatography. Hence, concentration of the enzyme can vary between batches and preliminary titrations are advised. Otherwise, the USER-compatible polymerase Phusion U is commercially available from Thermo Fisher Scientific.
18. We experience a higher frequency of correct transformants when gel purifying the PCRs prior to the USER assembly. However, for gBlock fragments it might prove sufficient with direct PCR purification using for example the GeneJET PCR Purification Kit.
19. Efficiency of the USER assembly depends on the number and sizes of the fragments for the assembly. If experiencing low efficiencies, consider plating all of the USER mixture on one plate.

Acknowledgments

This work was funded by grants from the Novo Nordisk Foundation to S.Y.L. and T.W. (NNF15OC0016226). The authors thank Günther Muth from the University of Tübingen for providing the pGM1190 plasmid.

References

- Berdy J (2005) Bioactive microbial metabolites – a personal view. *J Antibiot* 58(1):1–26
- Berdy J (2012) Thoughts and facts about antibiotics: where we are now and where we are heading. *J Antibiot* 65(8):385–395. doi:10.1038/Ja.2012.27
- Hwang KS, Kim HU, Charusanti P et al (2014) Systems biology and biotechnology of *Streptomyces* species for the production of secondary metabolites. *Biotechnol Adv* 32(2):255–268. doi:10.1016/J.Biotechadv.2013.10.008
- Blin K, Medema MH, Kazempour D et al (2013) antiSMASH 2.0-a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res* 41(W1):W204–W212. doi:10.1093/nar/gkt449
- Weber T, Blin K, Duddela S et al (2015) antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* 43(W1):W237–W243. doi:10.1093/nar/gkv437
- Weber T, Kim HU (2016) The secondary metabolite bioinformatics portal: computational tools to facilitate synthetic biology of secondary metabolite production. *Synthetic Syst Biotechnol* 1(2):69–79
- Ziemert N, Alanjary M, Weber T (2016) The evolution of genome mining in microbes – a review. *Nat Prod Rep* 33(8):988–1005. doi:10.1039/c6np00025h
- Weber T, Charusanti P, Musiol-Kroll EM et al (2015) Metabolic engineering of antibiotic factories: new tools for antibiotic production in actinomycetes. *Trends Biotechnol* 33(1):15–26. doi:10.1016/j.tibtech.2014.10.009
- Deveau H, Garneau JE, Moineau S (2010) CRISPR/Cas system and its role in phage-bacteria interactions. *Annu Rev Microbiol* 64:475–493. doi:10.1146/annurev.micro.112408.134123
- Koonin EV, Makarova KS (2013) CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biol* 10(5):679–686. doi:10.4161/rna.24022
- Barrangou R, Marraffini LA (2014) CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. *Mol Cell* 54(2):234–244. doi:10.1016/j.molcel.2014.03.011
- Abudayyeh OO, Gootenberg JS, Konermann S et al (2016) C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* 353(6299):aaf5573. doi:10.1126/science.aaf5573
- Makarova KS, Wolf YI, Alkhnbashi OS et al (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* 13(11):722–736. doi:10.1038/nrmicro3569
- Barrangou R, Fremaux C, Deveau H et al (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315(5819):1709–1712. doi:10.1126/science.1138140
- Bhaya D, Davison M, Barrangou R (2011) CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu Rev Genet* 45:273–297. doi:10.1146/annurev-genet-110410-132430
- Horvath P, Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and archaea. *Science* 327(5962):167–170. doi:10.1126/science.1179555
- Jinek M, Chylinski K, Fonfara I et al (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337(6096):816–821. doi:10.1126/science.1225829
- Deltcheva E, Chylinski K, Sharma CM et al (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471(7340):602–607. doi:10.1038/Nature09886
- Nishimasu H, Ran FA, Hsu PD et al (2014) Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* 156(5):935–949. doi:10.1016/j.cell.2014.02.001
- Sternberg SH, Redding S, Jinek M et al (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* 507(7490):62–67. doi:10.1038/Nature13011
- Iliakis G, Wang H, Perrault AR et al (2004) Mechanisms of DNA double strand break repair and chromosome aberration formation. *Cytogenet Genome Res* 104(1–4):14–20. doi:10.1159/000077461
- Kanaar R, Hoeijmakers JH, van Gent DC (1998) Molecular mechanisms of DNA double strand break repair. *Trends Cell Biol* 8(12):483–489
- Bassett AR, Tibbit C, Ponting CP et al (2013) Highly efficient targeted mutagenesis of *Drosophila* with the CRISPR/Cas9 system. *Cell Rep* 4(1):220–228. doi:10.1016/J.Celrep.2013.06.020
- DiCarlo JE, Norville JE, Mali P et al (2013) Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic Acids Res* 41(7):4336–4343
- Friedland AE, Tzur YB, Esvelt KM et al (2013) Heritable genome editing in *C. elegans* via a

- CRISPR-Cas9 system. *Nat Methods* 10 (8):741–743. doi:[10.1038/Nmeth.2532](https://doi.org/10.1038/Nmeth.2532)
26. Li DL, Qiu ZW, Shao YJ et al (2013) Heritable gene targeting in the mouse and rat using a CRISPR-Cas system. *Nat Biotechnol* 31 (8):681–683. doi:[10.1038/Nbt.2661](https://doi.org/10.1038/Nbt.2661)
 27. Mali P, Yang L, Esvelt KM et al (2013) RNA-guided human genome engineering via Cas9. *Science* 339(6121):823–826. doi:[10.1126/science.1232033](https://doi.org/10.1126/science.1232033)
 28. Ronda C, Pedersen LE, Hansen HG et al (2014) Accelerating genome editing in CHO cells using CRISPR Cas9 and CRISPy, a web-based target finding tool. *Biotechnol Bioeng* 111(8):1604–1616. doi:[10.1002/Bit.25233](https://doi.org/10.1002/Bit.25233)
 29. Wang HY, Yang H, Shivalila CS et al (2013) One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* 153(4):910–918. doi:[10.1016/J.Cell.2013.04.025](https://doi.org/10.1016/J.Cell.2013.04.025)
 30. Xie KB, Yang YN (2013) RNA-guided genome editing in plants using a CRISPRCas system. *Mol Plant* 6(6):1975–1983. doi:[10.1093/Mp/Sst119](https://doi.org/10.1093/Mp/Sst119)
 31. Yang DS, Xu J, Zhu TQ et al (2014) Effective gene targeting in rabbits using RNA-guided Cas9 nucleases. *J Mol Cell Biol* 6(1):97–99. doi:[10.1093/Jmcb/Mjt047](https://doi.org/10.1093/Jmcb/Mjt047)
 32. Lieber MR (1999) The biochemistry and biological significance of nonhomologous DNA end joining: an essential repair process in multicellular eukaryotes. *Genes Cells* 4(2):77–85
 33. Deriano L, Roth DB (2013) Modernizing the nonhomologous end-joining repertoire: alternative and classical NHEJ share the stage. *Annu Rev Genet* 47:433–455. doi:[10.1146/annurev-genet-110711-155540](https://doi.org/10.1146/annurev-genet-110711-155540)
 34. Aravind L, Koonin EV (2001) Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome Res* 11 (8):1365–1374. doi:[10.1101/gr.181001](https://doi.org/10.1101/gr.181001)
 35. Bowater R, Doherty AJ (2006) Making ends meet: repairing breaks in bacterial DNA by non-homologous end-joining. *PLoS Genet* 2 (2):e8. doi:[10.1371/journal.pgen.0020008](https://doi.org/10.1371/journal.pgen.0020008)
 36. Tong Y, Charusanti P, Zhang L et al (2015) CRISPR-Cas9 based engineering of actinomycetal genomes. *ACS Synth Biol* 4(9):1020–1029. doi:[10.1021/acssynbio.5b00038](https://doi.org/10.1021/acssynbio.5b00038)
 37. Tong Y, Weber, T, Lee, SY (2015) CRISPR/Cas9-based system. EU Patent EP15160126.7, 25 Mar 2015
 38. Qi LS, Larson MH, Gilbert LA et al (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* 152(5):1173–1183. doi:[10.1016/J.Cell.2013.02.022](https://doi.org/10.1016/J.Cell.2013.02.022)
 39. Czar MJ, Anderson JC, Bader JS et al (2009) Gene synthesis demystified. *Trends Biotechnol* 27(2):63–72. doi:[10.1016/j.tibtech.2008.10.007](https://doi.org/10.1016/j.tibtech.2008.10.007)
 40. Nour-Eldin HH, Hansen BG, Norholm MHH et al (2006) Advancing uracil-excision based cloning towards an ideal technique for cloning PCR fragments. *Nucleic Acids Res* 34(18):e122. doi:[10.1093/nar/gkl635](https://doi.org/10.1093/nar/gkl635)
 41. Bitinaite J, Rubino M, Varma KH et al (2007) USER (TM) friendly DNA engineering and cloning method by uracil excision. *Nucleic Acids Res* 35(6):1992–2002. doi:[10.1093/nar/gkm041](https://doi.org/10.1093/nar/gkm041)
 42. Genee HJ, Bonde MT, Bagger FO et al (2015) Software-supported USER cloning strategies for site-directed mutagenesis and DNA assembly. *ACS Synth Biol* 4(3):342–349. doi:[10.1021/sb500194z](https://doi.org/10.1021/sb500194z)
 43. Norholm MHH (2010) A mutant Pfu DNA polymerase designed for advanced uracil-excision DNA engineering. *BMC Biotechnol* 10:21. doi:[10.1186/1472-6750-10-21](https://doi.org/10.1186/1472-6750-10-21)
 44. Salomonsen BMUH, Halkier BA (2014) USER-derived cloning methods and their primer design. In: SLR V (ed) DNA cloning and assembly methods. *Methods in molecular biology*, vol 1116. Springer, New York
 45. Cavaleiro AM, Kim SH, Seppala S et al (2015) Accurate DNA assembly and genome engineering with optimized uracil excision cloning. *ACS Synth Biol* 4(9):1042–1046. doi:[10.1021/acssynbio.5b00113](https://doi.org/10.1021/acssynbio.5b00113)
 46. Cobb RE, Wang Y, Zhao H (2014) High-efficiency multiplex genome editing of *Streptomyces* species using an engineered CRISPR/Cas system. *ACS Synth Biol*. doi:[10.1021/sb500351f](https://doi.org/10.1021/sb500351f)
 47. Huang H, Zheng GS, Jiang WH et al (2015) One-step high-efficiency CRISPR/Cas9-mediated genome editing in *Streptomyces*. *Acta Biochim Biophys Sin (Shanghai)* 47(4). doi:[10.1093/abbs/gmv007](https://doi.org/10.1093/abbs/gmv007)
 48. Zeng H, Wen SS, Xu W et al (2015) Highly efficient editing of the actinorhodin polyketide chain length factor gene in *Streptomyces coelicolor* M145 using CRISPR/Cas9-CodA(sm) combined system. *Appl Microbiol Biotechnol* 99(24):10575–10585. doi:[10.1007/s00253-015-6931-4](https://doi.org/10.1007/s00253-015-6931-4)
 49. Kieser T, Bibb M, Buttner M et al (2000) *Practical Streptomyces genetics*. Norwich, UK, The John Innes Foundation
 50. Muth G, Nussbaumer B, Wohlleben W et al (1989) A vector system with temperature-

- sensitive replication for gene disruption and mutational cloning in streptomycetes. *Mol Gen Genet* 219(3):341–348. doi:[10.1007/Bf00259605](https://doi.org/10.1007/Bf00259605)
51. Xie SS, Shen B, Zhang CB et al (2014) sgRNA-cas9: a software package for designing CRISPR sgRNA and evaluating potential off-target cleavage sites. *PLoS One* 9(6). doi:[10.1371/journal.pone.0100448](https://doi.org/10.1371/journal.pone.0100448)
52. Blin K, Pedersen LE, Weber T et al (2016) CRISPy-web: an online resource to design sgRNAs for CRISPR applications. *Synthetic Syst Biotechnol* 1(2):4
53. Smithies O (2001) Forty years with homologous recombination. *Nat Med* 7(10): 1083–1086. doi:[10.1038/nm1001-1083](https://doi.org/10.1038/nm1001-1083)
54. Nour-Eldin HH, Geu-Flores F, Halkier BA (2010) USER cloning and USER fusion: the ideal cloning techniques for small and big laboratories. In: AG F-N (ed) *Plant secondary metabolism engineering*. *Methods in molecular biology*, vol 643. Springer, New York

Assembly and Multiplex Genome Integration of Metabolic Pathways in Yeast Using CasEMBLR

Tadas Jakočiūnas, Emil D. Jensen, Michael K. Jensen,
and Jay D. Keasling

Abstract

Genome integration is a vital step for implementing large biochemical pathways to build a stable microbial cell factory. Although traditional strain construction strategies are well established for the model organism *Saccharomyces cerevisiae*, recent advances in CRISPR/Cas9-mediated genome engineering allow much higher throughput and robustness in terms of strain construction. In this chapter, we describe CasEMBLR, a highly efficient and marker-free genome engineering method for one-step integration of in vivo assembled expression cassettes in multiple genomic sites simultaneously. CasEMBLR capitalizes on the CRISPR/Cas9 technology to generate double-strand breaks in genomic loci, thus prompting native homologous recombination (HR) machinery to integrate exogenously derived homology templates. As proof-of-principle for microbial cell factory development, CasEMBLR was used for one-step assembly and marker-free integration of the carotenoid pathway from 15 exogenously supplied DNA parts into three targeted genomic loci. As a second proof-of-principle, a total of ten DNA parts were assembled and integrated in two genomic loci to construct a tyrosine production strain, and at the same time knocking out two genes. This new method complements and improves the field of genome engineering in *S. cerevisiae* by providing a more flexible platform for rapid and precise strain building.

Key words Genome engineering, CRISPR/Cas9, Metabolic engineering, In vivo assembly, DNA assembly, CasEMBLR, Homologous recombination

1 Introduction

Budding yeast *S. cerevisiae* is a formidable host for biobased production of numerous chemicals and fuels, and also serves as an important model organism for functional genomics of eukaryotes [1–3]. With respect to biobased production, metabolic engineering in yeast is a rapidly growing field that calls for intelligent solutions to facilitate high-throughput strain building. Chromosomal integrations of heterologous genes encoding biosynthetic pathways are necessary to maintain stable production systems throughout

prolonged cultivations, whereas plasmid-based systems can be unstable [4]. Thus, chromosomal integrations are preferred, despite often increased strain construction time using this approach [5]. Chromosomal integrations of large biosynthetic pathways in yeast genomes have traditionally required sequential rounds of transformation in yeast. Though budding yeast is equipped with a high-fidelity DNA repair machinery to introduce heterologous DNA into its genome by homologous recombination, sequential transformations remains a challenge due to the low background efficiency of homologous recombination, and the limited number of selection markers needed to assist the selection of successful transformants from each round of transformation. These bottlenecks have made strain building in yeast a time-consuming process, and numerous approaches have been applied over the past 30 years to overcome these challenges [6–11]. The early studies on chromosomal integration dawned in the 1980s, where kb-sized homology arms flanked linear heterologous sequences [12, 13]. Since then, chromosomal integrations of heterologous sequences have been fine-tuned in native HR-based *in vivo* assemblies in multiple studies [6, 7].

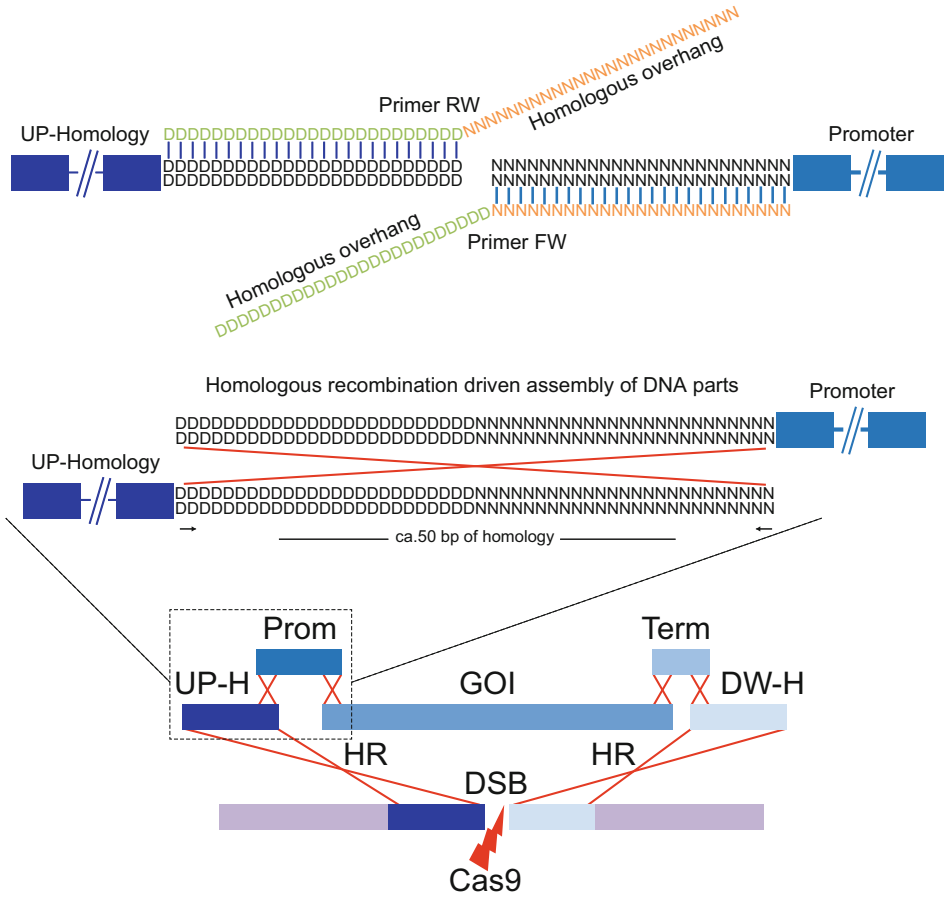
The key to improving the integration efficiency of heterologous DNA has been the co-integration of a marker-gene (often encoding *KIURA3*) along with the DNA sequence of interest, and then subsequently curing the strains for the marker [7]. In one study, the genome engineering method *Delitto perfetto* first introduces the counter-selectable (CORE) cassette containing *KIURA3* and *KanMX* into or nearby the sequence targeted for site-directed mutagenesis. The CORE cassette is later replaced by homologous recombination, using PCR-extended overlapping 80mers as donor templates [7]. The technique requires two transformations for a single mutagenesis with HR efficiency at 10^{-5} and counter-selection for *KIURA3* while testing for simultaneous loss of the *KanMX* cassette. The HR efficiency at 10^{-5} still falls within the range of previous reports relying on native HR for integration [11]. *Delitto perfetto* received an upgrade when introducing the *I-SceI* homing endonuclease site into the CORE cassette. Inducing the expression of the gene encoding *I-SceI* creates a double-strand break at the target site, hereby increasing integration efficiency to 5–20% [8]. This increase corresponds to 4000-fold higher integration efficiency, underscoring the importance of generating double-strand breaks at target sites for efficient genome engineering. *I-SceI* was later applied for assembling 10–15 parts into a single locus in yeast using the CATI procedure [9]. Here, marker-assisted integration efficiency of 95% was achieved, when the homing nuclease was induced. In sharp contrast, only 5% integration efficiency was reached when *I-SceI* was *not* induced. This result again highlights the need for generating a double-strand break to achieve efficient chromosomal integration.

One drawback of using homing endonucleases for genome engineering is the inherent dependence on the presence of genomic consensus endonuclease target sites. In order to avoid the need for introducing I-Sce target sites, and still harness the power of double-strand break at genomic integration sites for efficient genome engineering, CRISPR/Cas9 (Clustered-Regularly-Interspaced-Short-Palindromic-Repeats (CRISPR)-associated protein 9) has offered great improvements to the genome engineering workflow. The first study using Cas9 in yeast from DiCarlo et al. revealed how single editing events can be made without marker integrations. This new one-step procedure introduces a plasmid carrying a single gRNA expression cassette along with plasmid-borne Cas9 and edits the genome up to 100% efficiency, when co-transformed with a 90 bp double-stranded homology template as a donor [14]. This versatile tool has since been modified in various approaches [15–17]. In one study, Jakociūnas et al. further developed the system to allow for easy generation of multiple gRNA expression plasmids and thereby efficient multiplex genome engineering of up to five sites simultaneously [15]. Further, Bao et al. developed HI-CRISPR to target the genome in a marker-free manner [16]. In this system, the homology template was fused to gRNAs on the plasmid backbone, and the approach resulted in up to 100% integration efficiency at up to three simultaneously targeted sites. However, to reach this level, transformants were incubated in liquid culture for 6 days prior to plating.

As a next step for the improvement of yeast strain building we developed CasEMBLR, a tool for highly efficient and marker-free assembly and integration of multiple DNA components into genomic loci. Fast and efficient strain building depends on marker-free systems for multi-loci chromosomal integration. Thus, such a system should also be flexible in terms of choice of integration locus. CasEMBLR is a system harboring all these features. CasEMBLR is capable of efficiently integrating full metabolic pathways into yeast chromosomal DNA in a multiplex manner using *in vivo* DNA assembly, and it has significantly improved previously developed full pathway integration systems in yeast [6].

In this chapter, we describe the protocol for using CasEMBLR for *in vivo* assembly and integration of DNA parts. The CasEMBLR method relies on several steps: (1) each part for assembly is PCR amplified with specific primers designed to leave homologous overhangs to a neighboring part; (2) amplified parts are pooled together with gRNA expressing plasmid for targeting the integration site and transformed into the yeast strain expressing Cas9; (3) Cas9 introduces the double-strand break in the genomic site, which initiates DNA repair by homologous recombination, resulting in DNA part assembly and integration to the site of damage (Fig. 1a). As test-beds we show three examples for using CasEMBLR to constitute carotene gene expression cassettes and gene expression

a PCR amplification of DNA parts with specific primers containing 25 bp overhangs



b

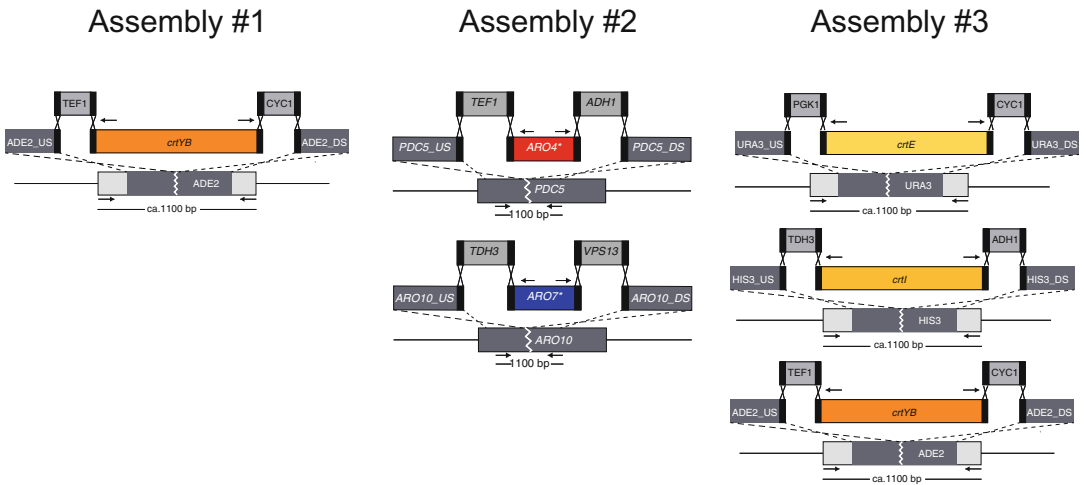


Fig. 1 Schematic overview of CasEMBLR method and assemblies performed. **(a)** Each assembly and integration event consists of five DNA parts: upstream homology region, promoter, gene of interest, terminator, and downstream homology part. Each part is amplified with specific primers leaving 25 bp overhangs, which allows a 50 bp homologous overlap between the parts. **(b)** Three assemblies that are described in the protocol are shown in detail

cassettes for high tyrosine production. Expression cassettes are assembled and integrated into one, two, or three sites simultaneously. Assembly #1—assembles and integrates expression cassette for carotene gene *crtYB* into ADE2 genomic site; Assembly #2—assembles and integrates expression cassette of mutated ARO4* to PDC5 genomic site and mutated ARO7* to ARO10 genomic site, simultaneously disrupting the PDC5 and ARO10 genes; Assembly #3—assembles and integrates expression cassette for carotene gene *crtE* into URA3 genomic site, *crtI* to HIS3 and *crtYB* to ADE2 simultaneously (Fig. 1b). Taken together, this method for in vivo assembly and multiplex integration of gene expression cassettes relieves expensive and time-consuming cloning procedures.

2 Materials

2.1 DNA

1. YIplac211-*crtYB*/*crtI*/*crtE*26 [18].
2. p414-TRP1-TEF1p-Cas9-CYC1t plasmid for Cas9 expression (obtained from Addgene: ID43802).
3. pCfB826 plasmid for amplification of mutated Aro4 and Aro7 variants [19].
4. PCR 2× Phusion High-Fidelity PCR master mix with HF buffer. (F-531S/L; Thermo Fisher Scientific). Any high fidelity polymerase can be used.
5. 3 M NaAc pH –5.2.
6. 96% EtOH.
7. Phusion U Hot Start DNA Polymerase (F-555S/L; Thermo Fisher Scientific).
8. USER enzyme (M5505S/L; New England Biolabs).

2.2 Transformation

1. 20% glucose solution.
2. SC-TRP liquid and solid media. Preparation of 1 L of liquid or solid media: 6.7 g of yeast nitrogen base without amino acids, 1.4 g of yeast synthetic drop-out media supplements without tryptophan (Sigma-Aldrich, Y1876), for solid media add 20 g of bacteriological agar, water to 900 ml. Sterilize by autoclavation and add 100 ml of 20% glucose.
3. SC-TRP-LEU liquid and solid media. Preparation of 1 L of liquid or solid media: 6.7 g of yeast nitrogen base without amino acids, 1.4 g of yeast synthetic drop-out media supplements without tryptophan and leucine (Sigma-Aldrich, Y0750), for solid media add 20 g of bacteriological agar, water to 900 ml. Sterilize by autoclavation and add 100 ml of 20% glucose.

4. 1 M Sorbitol. To prepare 1 L of 1 M Sorbitol add 182.17 g of Sorbitol (Sigma-Aldrich, S1000000) and fill the bottle up to 1 L with miliQ water. The solution has to be sterilized by filtration or other sterilization methods.
5. Transformation Enhancing Buffer (TEB). TEB contains: 100 mM Lithium acetate, 10 mM Dithiothreitol (DTT), 0.6 M Sorbitol, 10 mM Tris-HCl, pH -7.5. To Prepare 100 ml solution: mix 20 ml of miliQ water, 60 ml of 1 M Sorbitol, 10 ml of 1 M LiAc, 154 mg of DTT, and 10 ml of 100 mM Tris-HCl pH -7.5. TEB has to be stored in -20 °C.
6. TC-50 [20]: Yeast strain (CEN.PK111-27B – MATa leu2 trp1 + P414 – TRP1-TEF1p-Cas9-CYC1t) with Cas9 expressed from a single-copy centromeric plasmid .
7. Plasmid pTAJAK-96 [20]: Multi-copy, empty gRNA carrier plasmid containing leucine selection marker.
8. Plasmid pTAJAK-97 [20]: pTAJAK-96 carrying gRNA expression cassette for ADE2.
9. Plasmid pTAJAK-101 [20]: pTAJAK-96 carrying gRNA expression cassettes for PDC5 and ARO10.
10. Plasmid pTAJAK-105 [20]: pTAJAK-96 carrying gRNA expression cassettes for ADE2, URA3, and HIS3.
11. BioRad MicroPulser electroporator with settings “Fungi.”

2.3 Verification of Parts, Constructs, and Integrants

1. OneTaq Quick-Load 2xMaster mix (New England Biolabs, M0486S/L).
2. 1×TAE (Tris-Acetate Electrophoresis) buffer: 40 mM Tris-OH, 20 mM Acetic Acid, pH -7.8.
3. 1% Agarose gel: for 100 ml of the gel solution add 1 g of Agarose and 1×TAE buffer up to 100 ml. The volume of the gel solution depends on the electrophoresis equipment and the gel tray size. Heat the solution until agarose is completely dissolved, cool down to 60–70 °C, and pour the gel into the tray. Keep in room temperature until it is solid and ready to load.
4. DNA dye. Any DNA dye can be used for example EtBr or Safe-Red (Applied Biological Materials, G108-R).
5. DNA Gel loading dye (6×) (Thermo Fisher Scientific, R0611).

3 Methods

3.1 DNA Preparation for In Vivo Assembly

1. Yeast promoters (TDH3p, PGK1p, TEF1p), terminators (ADH1t, CYC1t, VPS13t, PRM9t) and homology sequences are amplified from genomic DNA isolated from *S. cerevisiae* CEN.PK113–7D. The carotenogenic genes crtYB (encoding a

bifunctional phytoene synthase and lycopene cyclase), crtI (phytoene desaturase), and crtE (heterologous geranylgeranyl pyrophosphate synthase) from the carotenoid producing yeast *Xanthophyllomyces dendrorhous* expressed from plasmid YIplac211-crtYB/crtI/crtE [18]. The ARO4*(K229L) and ARO7*(G141S) mutants were amplified from plasmid pCfB826 [19].

2. Primers for amplification of each part are listed in Table 1.
3. All necessary parts for assembly are amplified by Phusion High-Fidelity PCR master mix following the manufacturer's recommendations. Several reactions of the same part have to be prepared to have high enough concentration of each DNA part. We routinely prepare eight PCRs for each part.
4. Set up PCR reactions as follows: 25 μ l of 2 \times Phusion High-Fidelity PCR master mix, 50 ng of template DNA (genomic DNA or plasmid), 1.5 μ l of forward and 1.5 μ l of reverse primer, milliQ H₂O up to 50 μ l.
5. After PCR is finished, verify each reaction by loading 5 μ l of the reaction on 1% agarose gel and perform electrophoresis at 90 V for 30 min.
6. Mix together 4 pmol (*see Note 1*) of each part for the three assembly reactions that will be performed (Fig. 1; *see* Table 2 for the list of parts that has to be mixed together).
7. Concentrate mixed parts by ethanol precipitation (*see Note 2*) as follows: x μ l of DNA part + $3x$ μ l of 96% EtOH + 1/10 amount of total volume of 3 M NaAc (pH -5.6). Keep at -80 °C for 2 h or -20 °C overnight and centrifuge at 4 °C with a speed of 15,000–20,000 $\times g$ for 20 min. Pour out the supernatant and wash with 70% EtOH (if the DNA pellet is loose centrifuge for additional 5 min). Air-dry the pellet for 30–60 min and resuspend in 5 μ l of milliQ H₂O.
8. At this point parts can be stored at -20 °C for further use.

3.2 gRNA Selection and Plasmid Construction

1. CRISPy tool (http://staff.biosustain.dtu.dk/laeb/crispy_yeast/) is used to select most specific gRNAs for targeting ADE2, URA3, HIS3, ARO10, and PDC5. gRNA sequences are listed in Table 3 [20].
2. gRNA plasmids are constructed by amplifying gRNA expression cassettes (can be ordered from any DNA synthesis company) with Phusion U and USER cloning in carrier vector pTAJAK-96 with the LEU marker [15].

3.3 Yeast Transformation

1. Inoculate a single colony of yeast strain TC-50 (*see Note 3*) in 5 ml of SC-TRP medium and grow overnight at 30 °C shaking incubator (250 rpm).

Table 1
List of primers for DNA parts

| Primer name | Sequence | Description |
|--------------------|---|---|
| ASR_X001 | TCCTTAACCCCAACTGCACAGAA | Forward primer for URA3 upstream flanking sequence |
| ASR_X002 | GACCCCATTCCTTTGAAGGTACTTCTGACAATTCTGCTAACATCAAAAAG | Reverse primer for URA3 upstream flanking sequence with PGK1p overlap |
| ASR_X003 | CCTTTTGATGTTAGCAGAATTGTCAGGAAGTACCITTCAAAAGAAATGG | Forward primer for PGK1p with URA3 overlap |
| ASR_X004 | CTGTGAGGATGTTCCGCTAAATCCATTGTTTTATATTTTGTGTAATAAAGTAGATAA | Reverse primer for PGK1p with crtE overlap |
| ASR_X005 | TACTTTTTACAACAATAATAAAAACAATGGATTACGGCGAACATCCCTC | Forward primer for crtE with PGK1p overlap |
| ASR_X006 | TCCTTCCTTTTGGGTTAGAGCGGGATTCACAGAGGGATATCGGGCTAG | Reverse primer for crtE with tCYC1 overlap |
| ASR_X007 | AAAGCTAGCCGATATCCCTCTGTGAATCCGGCTCTAACCGGAAAAAGGA | Forward primer for tCYC1 with crtE overlap |
| ASR_X008 | GTC AACAGTACCCCTTAGTATATTCITTTCTCAAGCAAGGTTTTTCAGTATAATG | Reverse primer for tCYC1 with URA3 overlap |
| ASR_X009 | TTATACTGAAAACCTTGCTTGAGAAAAGAATATACTAAGGGTACTGTTGACATTG | Forward primer for URA3 downstream flanking sequence with tCYC1 overlap |
| ASR_X010 | CTGGCGAGGTAITGGATAGTT | Reverse primer for URA3 downstream flanking sequence |
| ASR_X133 | TGATGCATTACCTTGTTCATCTTCA | Forward primer for HIS3 upstream flanking sequence |
| ASR_X134 | CGAACTGAAAAAGCGTGTTTTTTATGCGGGGATTGCTCTCGGTC | Reverse primer for HIS3 upstream flanking sequence with TDH3p overlap |
| ASR_X136 | AAAGCTTGACCGGAGCAATCCCGCATATAAAAAACACGCTTTTTTCAGTTCCG | Forward primer for TDH3p with new HIS3 overlap |
| ASR_X018 | TATCCTGATCTTGTTCITTTTCCCATTTTGTGTTTATGTGTGTTTATTTCGA | Reverse primer for TDH3p with crtI overlap |
| ASR_X019 | GAATAAACACACATAAAACAAAACAAAATGGGAAAAGAACAAAGATCAGG | Forward primer for crtI with TDH3p overlap |

| | | |
|----------|--|--|
| ASR_X020 | AAATCATAAATCATAAGAAATTCGGCTCAGAAAAGCAAGAACACCAACG | Reverse primer for crtI with tADH1 overlap |
| ASR_X021 | ATCCGGTGGTGTCTTGCTTCTGAGCGAAATTTCTTATGATTTATGATTTTATTT | Forward primer for tADH1 with crtI overlap |
| ASR_X143 | TAGACGACCATCACACCACCTGAAGACATGCCGGTAGAGGTGTG | Reverse primer for tADH1 with HIS3 overlap |
| ASR_X145 | TATTTGACCACACCTCTACCGGCATGTCTTCAGTGGTGTGATGGTC | Forward primer for HIS3 downstream flanking sequence with tADH1 overlap |
| ASR_X146 | TTTTCTTTTTTCTATTACTCTTGGCCCT | Reverse primer for HIS3 downstream flanking sequence |
| ASR_X123 | CTGTAATGTAGTTCAGCCTTTGTTC | Forward primer for ADE2 upstream flanking sequence |
| ASR_X124 | AACATTTTGAAGCTATGGTGTGTGCCAAGGACAATATTTGTGACTTATGTT | Reverse primer for ADE2 upstream flanking sequence with TEF1p overlap |
| ASR_X125 | ACATAAGTCACAAAATATTGTCCCTGGCACACACCATAGCTTCAAA | Forward primer for TEF1p with ADE2 overlap |
| ASR_X032 | TCTGGTAATATGCGAGAGCCGTCATTTGTAAATTAACAACTTAGATTGCTATG | Reverse primer for TEF1p with crtYB overlap |
| ASR_X033 | AATCTAATCTAAGTTTTAAATTACAAATGACGGCTCTCGCATATTA | Forward primer for crtYB with TEF1p overlap |
| ASR_X034 | TCCTTCCCTTTTCGGTTAGAGCGGATTTACTIGCCCTTCCCAITCCGC | Reverse primer for crtYB with tCYC1 overlap |
| ASR_X035 | CATGAGCGGATGGGAAGGGCAGTAAATCCGGCTCTAACCCGAAAAAGGA | Forward primer for tCYC1 with crtYB overlap |
| ASR_X126 | TTCTTACCCAAATTGTAGAGACTATCTTCTCAAGCAAGGTTTTTCAGTATAATG | Reverse primer for tCYC1 with ADE2 overlap |
| ASR_X127 | TTATACTGAAAACCTTGCCTTGAGAAGATAGTCTTACAATTTGGGTAAGAAAAC | Forward primer for ADE2 downstream flanking sequence, overlap with tCYC1 |
| ASR_X128 | ACCACGTTAATGGCTCCCTT | Reverse primer for ADE2 downstream flanking sequence |
| ASR_X043 | TACCTAAACATCTATAACCTTCAAAAGTA | Forward primer for PDC5 upstream flanking sequence |
| ASR_X044 | AACATTTTGAAGCTATGGTGTGTGCATTTGGGTGCTACTATGCG | Reverse primer for PDC5 upstream flanking sequence with TEF1p overlap |

(continued)

Table 1
(continued)

| Primer name | Sequence | Description |
|-------------|---|---|
| ASR_X045 | TTCAGCGGCCATAGTAGCACCCCAATGCACACACCATAGCTTCAAA | Forward primer for TEF1p with PDC5 overlap |
| ASR_X046 | CAGCGAACATTTGGAGATTCACTCAATTTGTAATTAAAACTTAGAATT AGATTGCTATG | Reverse primer for TEF1p with Aro4pm overlap |
| ASR_X047 | AATCTAATCTAAGTTTTTAATTACAAATGAGTGAATCTCCAATGTTCCG | Forward primer for Aro4pm with TEF1p overlap |
| ASR_X048 | AAATCATAAATCATAAGAAATTCGCCATTTCTTGTAACTTCTC TTCITTTGTCCTG | Reverse primer for Aro4pm with tADH1 overlap |
| ASR_X049 | AAGAAGAGAAAGTTAACAAAGAAATAGCGGAATTTCTTATGATTTATGA TTTTTATTT | Forward primer for tADH1 with Aro4pm overlap |
| ASR_X050 | TTCCATTGGTTTCACAGTCGGGGCTGAGCGACCCTCATGCTATACCTG | Reverse primer for tADH1 with PDC5 overlap |
| ASR_X051 | TCTCAGGTATAGCATGAGGTCCGCTCAGCGCCGACTGTGAAACC | Forward primer for PDC5 downstream flanking sequence with tADH1 overlap |
| ASR_X052 | AACACCCAAAGATACGGTGG | Reverse primer for PDC5 downstream flanking sequence |
| ASR_X053 | TACCTATCAGTCAGTACGTCTCCAA | Forward primer for ARO10 upstream flanking sequence |
| ASR_X054 | CGAACTGAAAAAGCGTGTTTTTTATGTTGGTGAATTAAGCGCCTT | Reverse primer for ARO10 upstream flanking sequence with TDH3p overlap |
| ASR_X055 | CGTTC AAGGCGCTTAATTCACCAACATAAAAAACACGCG TTTTTTCAGTTCG | Forward primer for TDH3p with ARO10 overlap |
| ASR_X056 | CAGTTTCTGGTTTTGTGAAATCCATTTTGTGTTTGTATGTTATGTTG TGTTTATTCGA | Reverse primer for TDH3p with Aro7* overlap |

| | | |
|----------|--|---|
| ASR_X057 | GAATAAACACACATAAAACAAAATGGATTTCACAAAACCCAGAAAC | Forward primer for Aro7* with TDH3p overlap |
| ASR_X063 | AGCGGGTATATACTTTCATAIAGTGAITTACTIONTCCAACTTCTTAGCAAG | Reverse primer for Aro7* with tVPS13 overlap |
| ASR_X064 | CTTGCTAAGAAGGTTGGAAGAGTAATCACATATGAAAAGTATATACCCCGC | Forward primer for tVPS13 with Aro7* overlap |
| ASR_X065 | AITGGCTGTTTTAATAAACCCAGTATGAAGAAAAGTATAGCTCAACTATATTCCTT | Reverse primer for tVPS13 with ARO10 overlap |
| ASR_X066 | GAATATAGTTGAGCTATACTTTCTTCATACGTGGTTATTTAAACAGCCA | Forward primer for ARO10 downstream flanking sequence with tVPS13 overlap |
| ASR_X062 | TTAAGATATAAACTGGGCTAGTTTGC | Reverse primer for ARO10 downstream flanking sequence |

Table 2
Parts and their sizes for each assembly reaction

| | Assembly #1 | Size bp | Assembly #2 | Size bp | Assembly #3 | Size bp |
|----|-------------|---------|-------------|---------|-------------|---------|
| 1 | ADE2_US | 525 | PDC5_US | 525 | URA3_US | 525 |
| 2 | pTEF1 | 470 | pTEF1 | 470 | pPGK1 | 1034 |
| 3 | crtYB | 2072 | ARO4* | | crtE | 1181 |
| 4 | tCYC1 | 223 | tADH1 | 242 | tCYC1 | 223 |
| 5 | ADE_DS | 525 | PDC5_DS | 525 | URA3_DS | 525 |
| 6 | | | ARO10_US | 525 | HIS3_US | 525 |
| 7 | | | pTDH3 | 748 | pTDH3 | 748 |
| 8 | | | ARO7* | | crtI | 1827 |
| 9 | | | tVPS13 | | tADH1 | 242 |
| 10 | | | ARO10_DS | 525 | HIS3_DS | 525 |
| 11 | | | | | ADE2_US | 525 |
| 12 | | | | | pTEF1 | 470 |
| 13 | | | | | crtYB | 2072 |
| 14 | | | | | tCYC1 | 223 |
| 15 | | | | | ADE_DS | 525 |

Table 3
List of gRNAs

| Target site | Sequence |
|-------------|----------------------|
| ADE2 | AATTGTAGAGACTATCCACA |
| URA3 | TGCAAGGGCTCCCTATCTAC |
| HIS3_2 | ATCACACCACTGAAGACTGC |
| PDC5 | TTTCACAGTCGGCGCTCTAT |
| ARO10 | TTAATAACCACGTATGGCGT |

- Next day, measure the OD₆₀₀ and dilute it to OD₆₀₀ of 0.4 in 25 ml (enough for a single transformation) of fresh SC-TRP. Grow for approximately 4 h to reach the OD₆₀₀ of 1.4–1.6.
- Centrifuge the culture at 4000 rpm (2900 × *g*) for 3–5 min in room temperature and pour off the supernatant.
- Resuspend the cells in 10 ml of TEB. Incubate the cells in room temperature for 20–30 min.

5. Centrifuge at 4000 rpm ($2900 \times g$) for 3–5 min at 4 °C and pour off the supernatant.
6. Wash the cells with 20 ml ice-cold 1 M Sorbitol twice. Centrifuge at 4 °C and keep the cells on ice.
7. Centrifuge the cells at 4000 rpm ($2900 \times g$) for 1 min at 4 °C. Pour out the supernatant and resuspend in 50 μ l of ice-cold 1 M Sorbitol.
8. Add 5 μ l of mixed and precipitated DNA parts that are required to be assembled to the cells. Additionally, add 500 ng of single gRNA plasmid (pTAJAK-97) or 1 μ g of double gRNA plasmid, or 1.5 μ g of triple gRNA plasmid (*see Note 4*).
9. Keep on ice for 5 min.
10. Transfer cell-DNA mixture to a 2 mm ice-cold electroporation cuvette.
11. Electroporate the cells (*see Note 5*) and add 1 ml of 1 M Sorbitol immediately.
12. Transfer transformation mix from cuvette to a sterile 15 ml tube and incubate at 30 °C for 1 h without shaking.
13. Add 1 ml of SC-TRP-LEU medium and incubate with shaking at 30 °C for 1 h.
14. Centrifuge the cells at 4000 rpm ($2900 \times g$) for 3–5 min at room temperature and discard the supernatant.
15. Resuspend the cells in 50–100 μ l of SC-TRP-LEU and plate on SC-TRP-LEU plates.
16. Let the cells grow for 2–4 days at 30 °C.

3.4 Verification of Correct Assemblies and Integrations

1. To verify both correct assembly and integration of the transformed parts, multiplex colony PCR has to be performed.
2. Pick 12 colonies (or as desired) with pipette tip and resuspend in 2 μ l of miliQ water in a standard PCR tube.
3. Place the PCR tubes with resuspended cells in a thermo cycler for lysis.
4. Use the following program for cell lysis: 65 °C for 30 s, 8 °C for 30 s, 65 °C for 90 s, 97 °C for 180 s, 8 °C for 60 s, 65 °C for 180 s, 97 °C for 60 s, 65 °C for 60 s, 80 °C—hold.
5. Primers for verification of correct assemblies and integrations are listed in Table 4.
6. Prepare reaction for yeast colony PCR as described in Table 4.
7. Use the following PCR program: 94 °C—6 min,—40 \times (94 °C—35 s, 55 °C—45 s, 68 °C—2:30 min)—68 °C—10 min.—4 °C—hold. Optional, touch-down PCR step can be included in the program for amplification.
8. Run PCR reaction on 1% agarose gel. Correct band sizes are indicated in Table 4 (*see Note 6*).

Table 4
Genotyping setup to confirm correct parts assembly and integration

| crtI verification in HIS3 locus Component | crtE verification in URA3 locus Component | crtYB verification in ADE2 locus Component | ARO4* verification in PDC5 locus Component | ARO7* verification in ARO10 locus Component | Volume |
|--|--|---|---|--|-----------------------------|
| Water | Water | Water | Water | Water | 5 μ L |
| Primer: ASR_XI159 | Primer: ASR_X011 | Primer: ASR_XI29 | Primer: ASR_X067 | Primer: ASR_X071 | 1 μ L |
| Primer: ASR_X026 | Primer: ASR_X012 | Primer: ASR_X040 | Primer: ASR_X068 | Primer: ASR_X072 | 1 μ L |
| Primer: ASR_X027 | Primer: ASR_X013 | Primer: ASR_X041 | Primer: ASR_X069 | Primer: ASR_X073 | 0.5 μ L |
| Primer: ASR_XI60 | Primer: ASR_X014 | Primer: ASR_XI30 | Primer: ASR_X070 | Primer: ASR_X074 | 0.5 μ L |
| Lysed cells | Lysed cells | Lysed cells | Lysed cells | Lysed cells | 2 μ L |
| 2xOneTaq Master Mix with Quick-Load buffer | 2xOneTaq Master Mix with Quick-Load buffer | 2xOneTaq Master Mix with Quick-Load buffer | 2xOneTaq Master Mix with Quick-Load buffer | 2xOneTaq Master Mix with Quick-Load buffer | 10 μ L |
| <i>Total volume</i> | | | | | <i>20 μL</i> |
| Product sizes: 1.31 kb and 772 bp | Product sizes: 1.6 kb and 700 bp | Product sizes: 1.5 kb and 740 bp | Product sizes: 1.3 kb and 800 bp | Product sizes: 1.4 kb and 850 bp | |

3.5 Quantification of Coumaric Acid

1. PDC5::ARO4*; ARO10::ARO7* strain was constructed for the purpose to obtain higher titers of tyrosine. Tyrosine is a precursor for *p*-coumaric acid and can be produced directly from tyrosine by bacterial TAL (Tyrosine Ammonia-Lyase). Since *p*-coumaric acid is cell permeable and cannot be further metabolized by yeast cells, not like its precursor tyrosine, it makes it a good readout of the flux toward tyrosine.
2. HPLC is used to quantify *p*-coumaric acid produced by yeast cells. For the detailed procedure *see* Ref 20.

4 Notes

1. Lower concentrations can also be used depending on how many parts and in how many sites assemblies are integrated. Concentration as low as 0.5–1 pmol can be used if assembly is integrated in a single site. It is recommended to increase concentration of the parts to obtain high assembly/integration efficiency if assemblies are to be integrated in multiple sites.
2. Other concentration methods could be used. For transformation by electroporation the maximum volume of DNA depends on the electroporator or cuvettes used.
3. TC-50 is a CEN.PK111-27B yeast strain transformed with Cas9 expressing plasmid (P414—TRP1-TEF1p-Cas9-CYC1t, obtained from Addgene: ID43802).
4. Amount of gRNA plasmid needs to increase when increasing number of targets in the genome for gRNA-Cas9 to obtain enough colonies after transformation. For example for one target 500 ng of gRNA expression plasmid is used, for two targets 1 µg of plasmid, and for three targets 1.5 µg. The amount of gRNA plasmid will depend on each different case or different gRNA used.
5. The electroporation transformation method is used to obtain higher transformation efficiency. BioRad MicroPulser is used under settings for “Fungi” (standard parameters: 1.5 kV for 6 ms). Other transformation methods can also be used (Chemical transformation with LiAc). If other transformation methods are used, the amounts of parts and gRNA plasmid have to be adjusted experimentally.
6. Two bands on the gel of certain size (Table 5) will indicate a correct assembly and integration. If nothing was integrated in the desired locus, a single band of 1.1 kb will be observed. If integration occurred, but parts were not properly assembled, bands of different sizes may appear.

Table 5
Primer list for genotyping integrated assemblies

| | | |
|----------|-----------------------------|---|
| ASR_X011 | GAAGAAACATGAAATTGCCAG | Forward primer for URA3 upstream colony PCR |
| ASR_X014 | GTCTTATTGTTCTTGATTTGTGCC | Reverse primer for URA3 downstream colony PCR |
| ASR_X159 | TCTCGAGTTC AAGAGAAAAAAG | Forward primer for HIS3 upstream colony PCR |
| ASR_X160 | CACTTCAAATGGAACGCCAA | Reverse primer for HIS3 downstream colony PCR |
| ASR_X129 | CACTTCAAATGGAACGCCAA | Forward primer for ADE2 upstream colony PCR |
| ASR_X130 | AGACGGTAATACTAGATGCTGA | Reverse primer for ADE2 downstream colony PCR |
| ASR_X067 | GACAGGACAGAAAAAGTAATTACAAGA | Forward primer for PDC5 upstream colony PCR |
| ASR_X068 | ATGACAAAGAACTCTGTCGTC | Reverse primer for PDC5 upstream colony PCR |
| ASR_X069 | GCTGCTGCAGACAAAAGAA | Forward primer for PDC5 downstream colony PCR |
| ASR_X070 | GATGTCAAGGCTGAAACTAAGAAG | Reverse primer for PDC5 downstream colony PCR |
| ASR_X071 | CATATAAGTTGGGTTTGACTCATCAA | Forward primer for ARO10 upstream colony PCR |
| ASR_X072 | GGATGGTTTGCCCTCATAAACT | Reverse primer for ARO10 upstream colony PCR |
| ASR_X073 | TTATACCTATCACTAAGGAAGTTGA | Forward primer for ARO10 downstream colony PCR |
| ASR_X074 | TCCGGTATGTAATAGGTTAGTGG | Reverse primer for ARO10 downstream colony PCR |
| ASR_X012 | GAGTAACTCGAGTGGAAATGTC | Reverse primer for URA3 upstream colony PCR/crtE |
| ASR_X013 | TGGAAGCGATCCTGAAAAAG | Forward primer for crtE downstream colony PCR/URA3 |
| ASR_X026 | CCACGATGATAGCTGTGG | Reverse primer for HIS3 upstream colony PCR/crtI |
| ASR_X027 | GGTATACCTTTTGGTGTGTTG | Forward primer for crtI downstream colony PCR/HIS3 |
| ASR_X040 | AGCGGTAATAAGGTAGATAATGG | Reverse primer for ADE2 upstream colony PCR/crtYB |
| ASR_X041 | GAAAATCTTGAGTGTGGTCA | Forward primer for crtYB downstream colony PCR/ADE2 |

References

1. Li M, Borodina I (2015) Application of synthetic biology for production of chemicals in yeast *Saccharomyces cerevisiae*. *FEMS Yeast Res*. doi:[10.1111/1567-1364.12213](https://doi.org/10.1111/1567-1364.12213)
2. Giaever G, Chu AM, Ni L et al (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418:387–391. doi:[10.1038/nature00935](https://doi.org/10.1038/nature00935)
3. Lam FH, Ghaderi A, Fink GR, Stephanopoulos G (2014) Biofuels. Engineering alcohol tolerance in yeast. *Science* 346:71–75. doi:[10.1126/science.1257859](https://doi.org/10.1126/science.1257859)
4. Zhang Z, Moo-Young M, Chisti Y (1996) Plasmid stability in recombinant *Saccharomyces cerevisiae*. *Biotechnol Adv* 14:401–435. doi:[10.1016/S0734-9750\(96\)00033-X](https://doi.org/10.1016/S0734-9750(96)00033-X)
5. Özaydin B, Burd H, Lee TS, Keasling JD (2013) Carotenoid-based phenotypic screen of the yeast deletion collection reveals new genes with roles in isoprenoid production. *Metab Eng* 15:174–183. doi:[10.1016/j.ymben.2012.07.010](https://doi.org/10.1016/j.ymben.2012.07.010)
6. Shao Z, Zhao H, Zhao H (2009) DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res* 37:e16. doi:[10.1093/nar/gkn991](https://doi.org/10.1093/nar/gkn991)
7. Storici F, Lewis LK, Resnick MA (2001) In vivo site-directed mutagenesis using oligonucleotides. *Nat Biotechnol* 19:773–776. doi:[10.1038/90837](https://doi.org/10.1038/90837)
8. Storici F, Resnick MA (2003) Delitto perfetto targeted mutagenesis in yeast with oligonucleotides. *Genet Eng* 25:189–207
9. Kuijpers NGA, Chroumpi S, Vos T et al (2013) One-step assembly and targeted integration of multigene constructs assisted by the I-SceI meganuclease in *Saccharomyces cerevisiae*. *FEMS Yeast Res* 13:769–781. doi:[10.1111/1567-1364.12087](https://doi.org/10.1111/1567-1364.12087)
10. Siddiqui MS, Choksi A, Smolke CD (2014) A system for multilocus chromosomal integration and transformation-free selection marker rescue. *FEMS Yeast Res* 14:1171–1185. doi:[10.1111/1567-1364.12210](https://doi.org/10.1111/1567-1364.12210)
11. Wach A, Brachat A, Pöhlmann R, Philippsen P (1994) New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*. *Yeast* 10:1793–1808. doi:[10.1002/yea.320101310](https://doi.org/10.1002/yea.320101310)
12. Orr-Weaver TL, Szostak JW, Rothstein RJ (1981) Yeast transformation: a model system for the study of recombination. *Proc Natl Acad Sci U S A* 78:6354–6358. doi:[10.1073/pnas.78.10.6354](https://doi.org/10.1073/pnas.78.10.6354)
13. Szostak JW, Orr-Weaver TL, Rothstein RJ, Stahl FW (1983) The double-strand-break repair model for recombination. *Cell* 33:25–35. doi:[10.1016/0092-8674\(83\)90331-8](https://doi.org/10.1016/0092-8674(83)90331-8)
14. DiCarlo JE, Norville JE, Mali P et al (2013) Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic Acids Res*. doi:[10.1093/nar/gkt135](https://doi.org/10.1093/nar/gkt135)
15. Jakočiūnas T, Bonde I, Herrgård M et al (2015) Multiplex metabolic pathway engineering using CRISPR/Cas9 in *Saccharomyces cerevisiae*. *Metab Eng* 28:213–222. doi:[10.1016/j.ymben.2015.01.008](https://doi.org/10.1016/j.ymben.2015.01.008)
16. Bao Z, Xiao H, Liang J et al (2014) A homology integrated CRISPR-Cas (HI-CRISPR) system for one-step multi-gene disruptions in *Saccharomyces cerevisiae*. *ACS Synth Biol*. doi:[10.1021/sb500255k](https://doi.org/10.1021/sb500255k)
17. Ryan OW, Skerker JM, Maurer MJ et al (2014) Selection of chromosomal DNA libraries using a multiplex CRISPR system. *Elife*:e03703. doi:[10.7554/eLife.03703](https://doi.org/10.7554/eLife.03703)
18. Verwaal R, Wang J, Meijnen JP et al (2007) High-level production of beta-carotene in *Saccharomyces cerevisiae* by successive transformation with carotenogenic genes from *Xanthophyllomyces dendrorhous*. *Appl Environ Microbiol* 73:4342–4350. doi:[10.1128/AEM.02759-06](https://doi.org/10.1128/AEM.02759-06)
19. Rodriguez A, Kildegaard KR, Li M et al (2015) Establishment of a yeast platform strain for production of p-coumaric acid through metabolic engineering of aromatic amino acid biosynthesis. *Metab Eng* 31:181–188. doi:[10.1016/j.ymben.2015.08.003](https://doi.org/10.1016/j.ymben.2015.08.003)
20. Jakočiūnas T, Rajkumar AS, Zhang J et al (2015) CasEMBLR: Cas9-facilitated multilocus genomic integration of in vivo assembled DNA parts in *Saccharomyces cerevisiae*. *ACS Synth Biol*. doi:[10.1021/acssynbio.5b00007](https://doi.org/10.1021/acssynbio.5b00007)

A Modified Gibson Assembly Method for Cloning Large DNA Fragments with High GC Contents

Lei Li, Weihong Jiang, and Yinhua Lu

Abstract

Gibson one-step, isothermal assembly method (Gibson assembly) can be used to efficiently assemble large DNA molecules by in vitro recombination involving a 5'-exonuclease, a DNA polymerase and a DNA ligase. In the past few years, this robust DNA assembly method has been widely applied to seamlessly construct genes, genetic pathways and even entire genomes. Here, we expand this method to clone large DNA fragments with high GC contents, such as antibiotic biosynthetic gene clusters from *Streptomyces*. Due to the low isothermal condition (50 °C) in the Gibson reaction system, the complementary overlaps with high GC contents are proposed to easily form mismatched linker pairings, which leads to low assembly efficiencies mainly due to vector self-ligation. So, we modified this classic method by the following two steps. First, a pair of universal terminal single-stranded DNA overhangs with high AT contents are added to the ends of the BAC vector. Second, two restriction enzyme sites are introduced into the respective sides of the designed overlaps to achieve the hierarchical assembly of large DNA molecules. The optimized Gibson assembly method facilitates fast acquisition of large DNA fragments with high GC contents from *Streptomyces*.

Key words *Streptomyces*, Gibson assembly, High GC content, Antibiotic biosynthetic gene cluster

1 Introduction

Streptomyces species are high GC Gram-positive bacteria, well known for their capability to produce a vast number of secondary metabolites, which exhibit highly diverse chemical structures and important biological activities [1, 2]. Secondary metabolites are generally produced by the compact biosynthetic gene clusters (in size from several kilobases to more than 100 kb). To discover novel secondary metabolites or improve the production levels of known important metabolites, it is of great importance to develop new tools for manipulating secondary metabolites biosynthetic gene clusters, such as cloning, editing, amplification, and deletion. Here, we present a modified Gibson assembly method to clone antibiotic biosynthetic gene clusters from *Streptomyces*.

Based on in vitro homologous recombination, Gibson isothermal assembly method exhibits high simplicity and speed to assemble large DNA constructs (up to 900 kb) and permits sequence-independent, one-pot assembly of multiple DNA fragments. Three enzymes, including T5 exonuclease, Phusion DNA polymerase, and Taq DNA ligase, are used to perform the ligation between the overlapping DNA molecules in the Gibson reaction system. Terminal single-stranded DNA overhangs are firstly produced by T5 exonuclease, then repaired and sealed by the latter two enzymes [3, 4]. However, we showed that this method couldn't be effectively used to assemble DNA fragments with high GC contents (>70%) from *Streptomyces* mainly due to vector self-ligation [5]. It is proposed that mismatched linker pairings would easily be generated between the overhangs with high GC contents due to low reaction temperature (50 °C). In fact, Casini et al. also reported that assembly efficiencies would decline along with increases in GC contents of DNA fragments [6].

To address this problem, we modified the classical Gibson assembly method. First, two universal terminal single-stranded DNA overhangs with high AT contents are added to the ends of the BAC vector, which significantly decreased the rate of vector self-ligation. This novel design also makes the linearized BAC vector eligible for repeated uses in independent assemblies. In the classical Gibson method, different linearized BAC vectors has to be prepared in each reaction for assembling different DNA fragments. Second, we introduced two restriction enzyme sites (*Nde* I/*Nhe* I) into the respective sides of the designed overhangs to achieve the hierarchical and seamless assembly of large DNA molecules. Thus, the left, middle, and right assembled fragments in the next assembly process can be obtained by digestion with *Nhe* I, *Nhe* I/*Nde* I, and *Nde* I, respectively (Fig. 1).

2 Materials

2.1 Gibson Assembly

1. 5× isothermal reaction buffer (600 µL): 300 µL of 1 M Tris-HCl, pH 7.5, 15 µL of 2 M MgCl₂, 24 µL of 100 mM dNTP (Generay), 30 µL of 1 M DTT (Sigma), 0.15 g of PEG-8000 (Sigma), 30 µL of 100 mM NAD (Sigma) and 700 µL of double-distilled water. Store at -20 °C.
2. Gibson assembly master mixture (1.2 mL): 320 µL of 5× isothermal reaction buffer, 0.64 µL of 10 U/µL T5 exonucleases (Epicentre), 20 µL of 2 U/µL Phusion polymerase (NEB), 160 µL of 40 U/µL Taq ligase (NEB) and 700 µL of double-distilled water. Store at -20 °C (*see Note 1*).

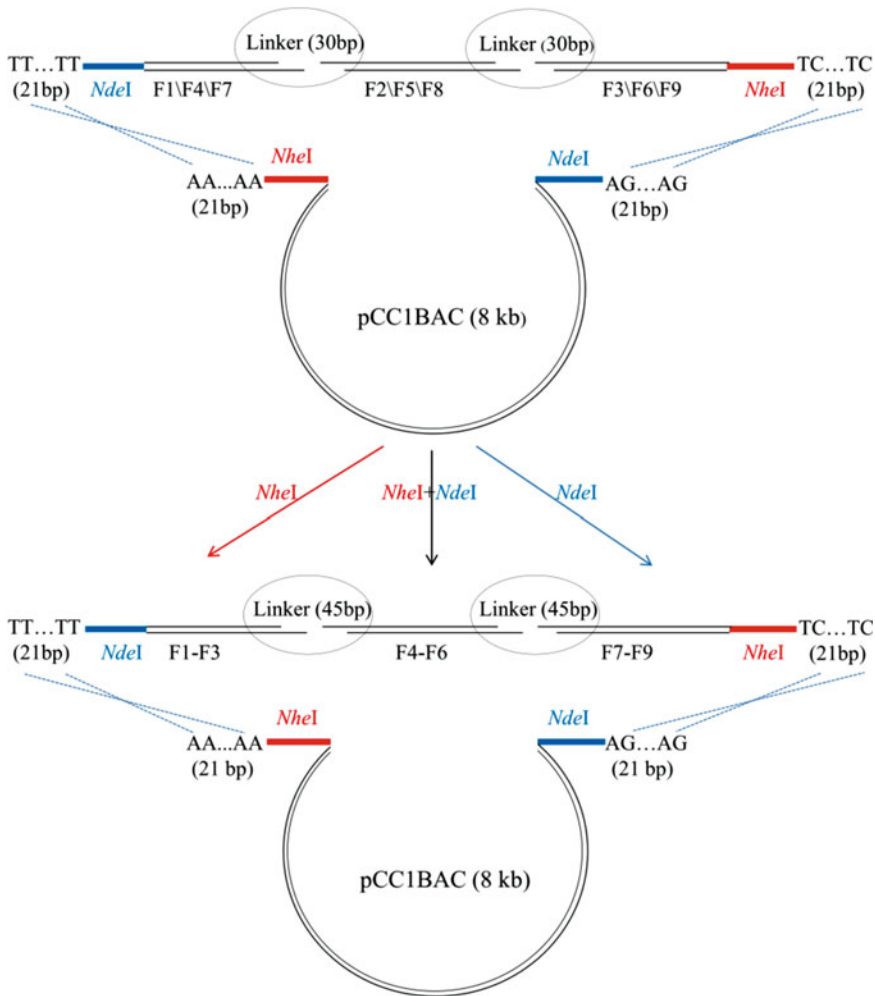


Fig. 1 The modified Gibson assembly method for assembling large DNA fragments with high GC contents from *Streptomyces*. In the first level, three DNA fragments (each with a size of 5 kb) are assembled simultaneously using the BAC vector (pCC1BAC). The complementary overlaps between two DNA inserts are 30 bp in length. In the second level, the left, middle, and right assembled products (F1-F3, F4-F6, and F7-F9, respectively) from the first level are digested by *Nhe* I, *Nhe* I/*Nde* I, and *Nde* I for the next level of assembly, respectively. The overlaps between two DNA inserts are 45 bp at the second level

2.2 BAC and DNA Purification

1. Bacterial artificial chromosome (BAC) vector pCC1BAC (Epicentre) and its derived recombinants are prepared using an E.Z. N.A.[®] BAC/PAC DNA Kit (Omega Bio-Tek).
2. PCR products are purified using an AxyPrep DNA Gel Extraction Kit (Axygen).
3. Restrictive enzymatic digestion products are purified using an Agarose Gel DNA Extraction Kit (Roche).
4. Isolation of the *Streptomyces pristinaespiralis* genomic DNA is performed as described by Kieser et al. (2000) [7].

2.3 Medium

1. Liquid LB medium: 1% NaCl, 1% Peptone, and 0.5% Yeast extract.
2. Solid LB medium: 1% NaCl, 1% Peptone, 0.5% Yeast extract, and 2% Agar.
3. SOB medium: 0.05% NaCl, 0.0186% KCl, 2% Peptone, and 0.5% Yeast extract.
4. SOC medium: SOB + 20 mM Glucose.

3 Methods

3.1 Preparation of DNA Fragments and the Linearized BAC Vector

1. The universal linearized BAC vector is produced by PCR amplification using a primer pair BAC-fw and BAC-rev (the sequences are presented below), followed by digestion with *Dpn* I for 2 h at 37 °C to remove the circular template. The universal overlaps (bold and italic characters in the primer sequences) and restriction enzyme sites (*Nde* I and *Nhe* I) (*see Note 2*) are introduced into the primers.

BAC-fw: 5'-***AGAGAGAGAGAGAGAGAGAGACATATG***-
 GATCCTCTAGAGTCGACCT-3' BAC-rev: 5'-
AAAAAAAAAAAAAAAAAAAAAAAAAGCTAGCCGGG-
 TACCGAGCTCGAATT-3'

2. DNA fragments in the first level of DNA assembly are produced by PCR amplification. The complementary overlaps between DNA fragments and restriction enzyme sites are introduced by primers. For example, the sequence 5'-TT...TT-3' (21 bp) and the *Nde* I restriction site are added into the F1\F4\F7 fragments, and the sequence 5'-TC...TC-3' (21 bp) and the *Nhe* I restriction site are introduced into the fragments of F3\F6\F9 (*see Note 3*).
3. DNA fragments in the latter levels are generated by digestion with appropriate restriction enzymes; for example, *Nde* I for the left assembled fragments, *Nde* I and *Nhe* I for the middle, and *Nhe* I for the right. After digestion, the assembled DNA fragments are isolated by 1% agarose gel electrophoresis and then purified using an Agarose Gel DNA Extraction Kit (*see Fig. 2 and Note 4*).

3.2 Preparation of *E. coli* EPI300 Electrocomponent Cells with High Transformation Efficiency

1. Streak out frozen glycerol stock of EPI300 onto an LB plate (without antibiotics). Grow overnight at 37 °C.
2. Pick a single colony into SOB medium and cultivated for 12 h at 37 °C, 200 rpm.
3. Transfer 200 μL of the culture into 200 mL of SOB medium in a 500 mL shake flask at 37 °C, 200 rpm. When the cell density (OD₅₅₀) reaches approximately 0.5, put the cells on ice immediately.

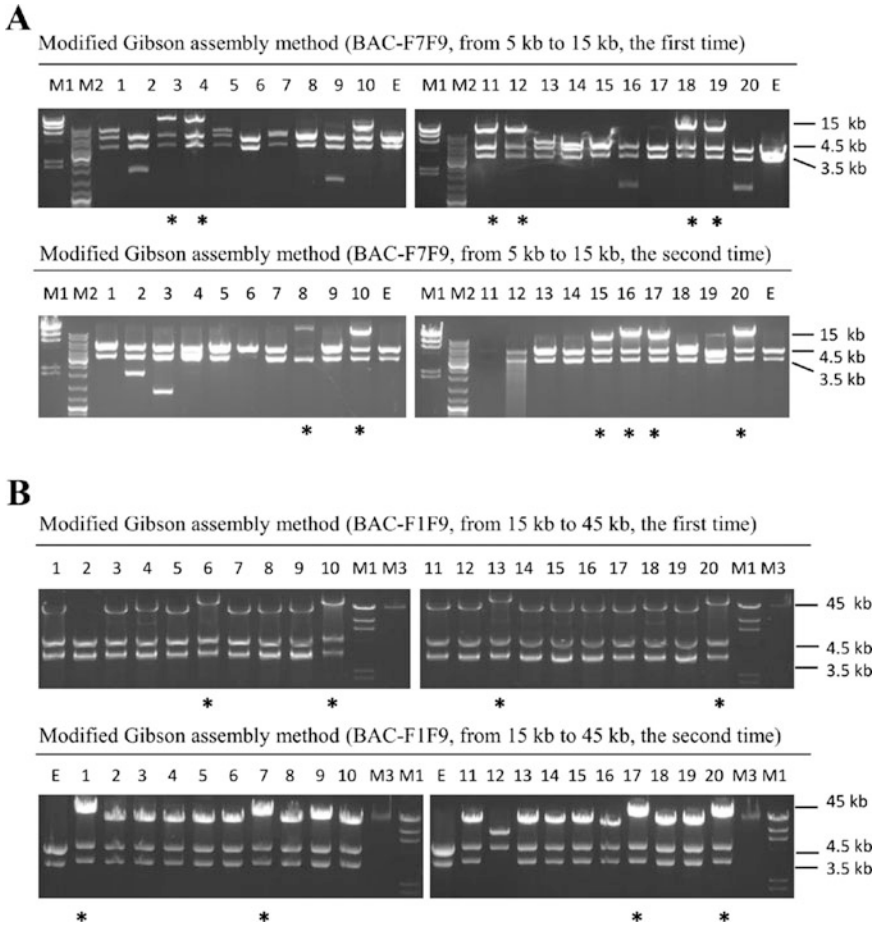


Fig. 2 Assembly efficiencies of the modified Gibson assembly method. *Nde* I restriction analysis of the recombinant BAC plasmids that were isolated from 20 randomly picked clones of the first-level assembly from 5 kb to 15 kb (**a**) and the second-level assembly from 15 kb to 45 kb (**b**). M1, M2 and M3 represent the 1 kb ladder, λ *Hind* III DNA ladder and λ DNA (NEB), respectively. E represents the pCC1BAC control vector with universal overlaps. Asterisk (*) indicates the correct assembly

4. Harvest the cells by centrifugation at $3000 \times g$ for 10 min at 4°C and wash the cells with ice-cold 10% glycerol three times.
5. Resuspend the pellet in 1 mL of ice-cold 10% glycerol. The final OD_{550} of the resuspended cells should be 200–250. Store frozen cells at -80°C .

3.3 Gibson Assembly

1. Mix equimolar DNA fragments (normally three fragments) with 80 ng of the linearized BAC vector. For different sizes of DNA segments, proportional amounts of DNA are added; for example, 50 ng of each 5 kb DNA segments, 150 ng of each 15 kb DNA segments, and 450 ng of each 45 kb DNA segments. Add double-distilled water to a total volume of 5 μL .

2. Add 5 μL of the mixture of the linearized BAC vector and DNA fragments to 15 μL of the Gibson assembly master mixture. Incubate for 1 h at 50 °C.
3. Transform 1 μL of the assembly mixture into 30 μL *E. coli* EPI300 electrocompetent cells in a 1 mm cuvette at 1, 200 V, 25 mF and 200 Ω .
4. Recover the cells at 37 °C, 200 rpm for 2 h in 1 mL SOC medium and then plate onto LB plates containing 12.5 $\mu\text{g}/\text{mL}$ chloramphenicol. Incubate at 37 °C, 200 rpm for 24 h.
5. Pick 10 transformants and then grow in 5 mL LB medium with 12.5 $\mu\text{g}/\text{mL}$ chloramphenicol overnight at 37 °C, 200 rpm.
6. Isolate the recombinant BAC plasmids from these cells using an E.Z.N.A.[®] BAC/PAC DNA Kit.
7. Digest the recombinant BAC vectors (0.5–0.8 μg) with *Nde* I at 37 °C for 2 h and perform 1% agarose gel-electrophoresis to identify the correct recombinants. The empty vector pCC1BAC is used as a negative control.
8. Enrich the correct recombinant BAC plasmids. Digest the plasmids BAC-F1F3, BAC-F4F6, and BAC-F7F9 with the restriction enzymes *Nhe* I, *Nhe* I/*Nde* I, and *Nde* I, respectively and then purify the corresponding assembled DNA fragments for the next level of DNA assembly from **step 1** (Figs. 1 and 2).

4 Notes

1. The Gibson assembly master mixture can be stored at –20 °C for at least 1 year and the enzymes remain active following at least 10 freeze-thaw cycles.
2. Restriction enzyme sites should be adjusted according to different antibiotic biosynthetic gene clusters. The chosen sites can't exist in the cloned gene cluster. For DNA fragments with high GC contents from *Streptomyces*, the following sites are recommended: *Nde* I, *Nhe* I, *Xba* I, *Spe* I, *Pme* I, *Swa* I, or *Pac* I.
3. Along with increases in the assembly sizes, longer complementary overlaps between DNA fragments should be designed to achieve high assembly efficiencies. For example, 30 bp for 5 kb, 45 bp for 15 kb and 60 bp for 45 kb. We also suggest that the universal overlaps between DNA fragments and the vector pCC1BAC should be appropriately extended when the size of assembled DNA fragments is beyond 50 kb.
4. When the Agarose Gel DNA Extraction Kit is used to purify restrictive enzymatic digestion products, the experiments should be carried out moderately and carefully to avoid the breakage of linearized DNA fragments.

Acknowledgement

This work was financed by the grants from National Natural Science Foundation of China (31630003, 31421061, 31370081, and 31570072) and the National Basic Research Program of China (2012CB721103).

References

1. Newman DJ, Cragg GM (2007) Natural products as sources of new drugs over the last 25 years. *J Nat Prod* 70(3):461–477
2. Bode HB, Muller R (2005) The impact of bacterial genomics on natural product research. *Angew Chem Int Edit* 44(42):6828–6846
3. Gibson DG, Young L, Chuang RY et al (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6(5):343–347
4. Gibson DG, Benders GA, Andrews-Pfannkoch C et al (2008) Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* 319(5867):1215–1220
5. Li L, Zhao YW, Ruan LJ et al (2015) A stepwise increase in pristinamycin II biosynthesis by *Streptomyces pristinaespiralis* through combinatorial metabolic engineering. *Metab Eng* 29:12–25
6. Casini A, MacDonald JT, De Jonghe J et al (2014) One-pot DNA construction for synthetic biology: the modular overlap-directed assembly with linkers (MODAL) strategy. *Nucleic Acids Res* 42:e7
7. Kieser T, Bibb MJ, Butter MJ et al (2000) Practical *Streptomyces* genetics. The John Innes Foundation, Norwich

Coupling Yeast Golden Gate and VEGAS for Efficient Assembly of the Violacein Pathway in *Saccharomyces cerevisiae*

James Chuang, Jef D. Boeke, and Leslie A. Mitchell

Abstract

The ability to express non-native pathways in genetically tractable model systems is important for fields such as synthetic biology, genetics, and metabolic engineering. Here we describe a modular and hierarchical strategy to assemble multigene pathways for expression in *S. cerevisiae*. First, discrete promoter, coding sequence, and terminator parts are assembled in vitro into Transcription Units (TUs) flanked by adapter sequences using “yeast Golden Gate” (yGG), a type IIS restriction enzyme-dependent cloning strategy. Next, harnessing the natural capacity of *S. cerevisiae* for homologous recombination, TUs are assembled into pathways and expressed using the “Versatile Genetic Assembly System” (VEGAS) in yeast. Coupling transcription units constructed by yGG with VEGAS assembly is a generic and flexible workflow to achieve pathway expression in *S. cerevisiae*. This protocol describes assembly of a five TU pathway for yeast production of violacein, a pigment derived from *Chromobacterium violaceum*.

Key words Metabolic engineering, *Saccharomyces cerevisiae*, Yeast Golden Gate (yGG), Versatile Genetic Assembly System (VEGAS), Violacein, Hierarchical DNA assembly

1 Introduction

Highly engineered, recombinant organisms have the capacity to serve as tiny cellular factories for the production of high value compounds. With expanding knowledge of an increasingly broad range of microbial hosts, plus new molecular engineering tools to redirect native pathways and express non-native genes, we are moving towards an industrial biotechnological future in which any organic molecule of interest can be made in cells with relative ease.

Saccharomyces cerevisiae, a model organism highly amenable to genetic manipulation and an established industrial workhorse, is a favored chassis for eukaryotic metabolic engineering. A diversity of small molecules have been successfully produced in yeast, ranging from isoprenoids [1], to fatty acids [2], to flavonoids [3], and

beyond (reviewed in [4]). Underlying the success of yeast as a chassis is a rapid design–build–test cycle facilitated by the enormous trove of systems level information available for yeast [5] plus the inherent capacity of *S. cerevisiae* to perform homologous recombination. Thus, engineering the *S. cerevisiae* genome to redirect the flux of native biosynthetic pathways together with the introduction of non-native genes for expression is readily achievable.

Here we detail methods for modular and efficient assembly of non-native genes into pathways for expression in *S. cerevisiae*. The assembly of genes, or transcription units (TUs) is achieved using a standardized version of Golden Gate cloning we call “yeast Golden Gate” (yGG) [6]. Subsequently, TUs are assembled into pathways for expression in yeast using “VEGAS,” a Versatile Genetic Assembly System [7]. In brief, yGG specifies assembly (from left-to-right) of a left VEGAS adapter (LVA) sequence, a promoter (PRO), a coding sequence (CDS), a terminator (TER), and a right VEGAS adapter (RVA) sequence into an acceptor vector (Fig. 1a). The PRO and TER sequences are abstracted from the *S. cerevisiae* genome using defined rules [6] and all parts in yGG reactions are flanked by type IIS restriction enzyme sites that expose designer overhangs and enable directional assembly (Table 1). The VEGAS adapters then serve as templates to generate terminal homology between adjacent parts for overlap-directed assembly by homologous recombination in yeast (Fig. 1b). The yGG-VEGAS strategy permits facile assembly of multigene pathways in yeast. Importantly, modularity of the yGG step supports fine-tuning of gene expression through the use of combinatorial libraries of PRO and TER parts as we have previously demonstrated for carotenoid production [7]. Here we provide a detailed protocol outlining the yGG-VEGAS method to achieve production of violacein in *S. cerevisiae*.

2 Materials

Solutions can be prepared using deionized water. All liquids should be sterilized by autoclaving and stored at room temperature unless otherwise noted. For long-term storage, DNA stocks should be stored at $-20\text{ }^{\circ}\text{C}$. Plasmids required to assemble and express the violacein pathway in yeast may be requested from the lab of Jef Boeke at New York University Medical Center.

2.1 Yeast Golden Gate (yGG) Assembly Reagents

1. Yeast Golden Gate acceptor vector: pJC120.
2. Donor DNA parts for yGG: violacein assembly parts are listed in Table 2 with plasmid numbers indicated (*see Note 1*).
3. Spectrophotometer for nucleic acid quantification.

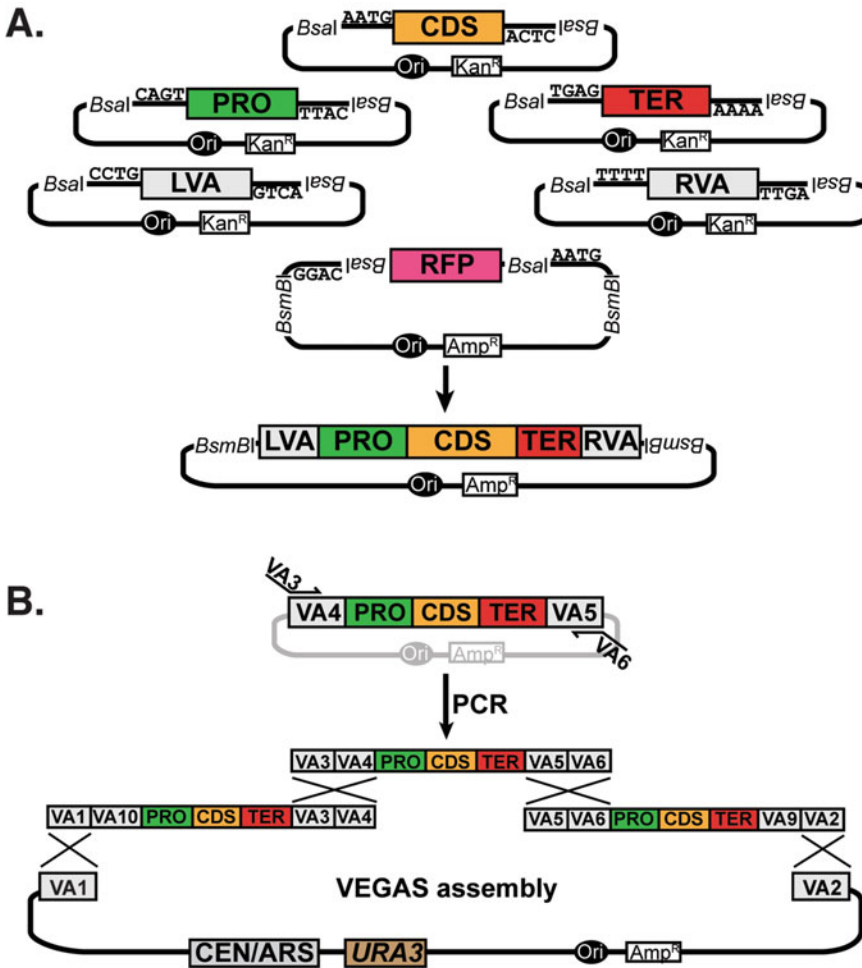


Fig. 1 Schematic outlining yGG coupled to VEGAS to assemble pathways for expression in *S. cerevisiae*. (a) Yeast Golden Gate (yGG) to assemble yeast transcription units (TUs). Donor parts include left and right VEGAS adapters (LVA, RVA, respectively), coding sequence (CDS), promoter (PRO), and terminator (TER). Donor parts are encoded on kanamycin resistance vectors and parts are flanked by inwardly pointing *Bsal* sites. Designer overhangs exposed following *Bsal* digestion are indicated. The yGG acceptor vector, carrying ampicillin resistance, encodes an RFP gene for expression in *E. coli*. In a one-pot Golden Gate reaction, the TU (PRO-CDS-TER) flanked by LVA and RVA sequences is assembled. Assembled constructs can be distinguished from unmodified parental acceptor following transformation of yGG product into *E. coli* by red–white screening. *BsmBI* restriction digestion can release the assembled TU from the yGG acceptor vector. (b) PCR-mediated VEGAS for assembly of pathways for expression in *S. cerevisiae*. TUs derived from yGG assembly are amplified using primers encoding terminal homology between adjacent parts. Cotransformation into yeast together with a linearized VEGAS assembly vector enables homologous recombination in yeast and pathway assembly

4. Bovine Serum Albumin (BSA): prepare a 1 mg/mL working solution by diluting a 20 mg/mL stock 1:20 with water.
5. 10× T4 DNA ligase reaction buffer: 500 mM Tris–HCl pH 7.5, 100 mM MgCl₂, 100 mM dithiothreitol, 10 mM ATP. Aliquot and store at –20 °C (see Note 2).

Table 1
Overhangs for yGG assembly

| Part | 5' overhang | 3' overhang |
|---------------------------|-------------|-------------|
| Left VEGAS adapter (LVA) | CCTG | CAGT |
| Promoter | CAGT | AATG |
| CDS | AATG | TGAG |
| Terminator | TGAG | TTTT |
| Right VEGAS adapter (RVA) | TTTT | AACT |

Table 2
Parts list for violacein assembly

| TU order (left to right) | LVA | PRO | CDS | TER | RVA | Assembled TU length (bp) |
|--------------------------------|------------------|-------------------|--------------------------------|--------------------|------------------|-----------------------------|
| 1 | VA7 (pLM400) | pTDH3 (pJC020) | vioA (pJC074) | ttACS2 (pJC006) | VA3 (pJC130) | 2087 |
| 2 | VA8 (pJC143) | pPGK1 (pJC021) | vioB (pJC089) | ttENO2 (pJC003) | VA4 (pJC144) | 3827 |
| 3 | VA9 (pJC145) | n/a | kanMX ^a (pJC065) | n/a | VA5 (pJC146) | 1529 |
| 4 | VA10 (pJC165) | pACT1 (pJC018) | vioC (pJC076) | ttASC1 (pJC005) | VA6 (pJC147) | 2120 |
| 5 | VA11 (pJC157) | pRPS2 (pJC019) | vioD (pJC077) | ttCIT1 (pJC022) | VA12 (pJC154) | 1952 |
| 6 | VA16 (pJC149) | pZEO1 (pJC114) | vioE (pJC078) | ttFUM1 (pJC004) | VA5 (pJC146) | 1314 |

^aSee Note 11

6. T4 DNA ligase, 600 U/μL.
7. Restriction endonuclease *Bsa*I (see Note 3).
8. Restriction endonuclease *Bsm*BI.
9. PCR tubes (0.2 mL).
10. Microcentrifuge tubes (1.5 mL).

2.2 *E. coli* Growth and Transformation Reagents

1. Chemically competent DH5α *E. coli*: store at −80 °C (see Note 4).
2. Carbenicillin stock, 75 mg/mL: dissolve 750 mg of carbenicillin disodium salt in a final volume of 10 mL and filter-sterilize (see Note 5). Aliquot and store at −20 °C.

3. Kanamycin stock, 50 mg/mL: dissolve 500 mg of kanamycin sulfate in a final volume of 10 mL and filter-sterilize. Aliquot and store at -20°C .
4. LB liquid medium: in a final volume of 1 L, dissolve 10 g of bacto tryptone, 5 g of yeast extract, and 10 g of sodium chloride. Autoclave immediately.
5. LB liquid medium with carbenicillin: supplement 500 mL of LB liquid medium with 500 μL of 75 mg/mL carbenicillin. Mix well and store at 4°C .
6. LB liquid medium with kanamycin: supplement 500 mL of LB liquid medium with 500 μL of 50 mg/mL kanamycin. Mix well and store at 4°C .
7. Drum rotator for culture tubes.
8. 30°C incubator.
9. Sterile culture tubes.
10. LB agar plates with kanamycin: in a final volume of 1 L, dissolve 10 g of bacto tryptone, 5 g of yeast extract, 10 g of sodium chloride; once dissolved add 20 g of bacto agar and autoclave immediately. Cool to $\sim 50^{\circ}\text{C}$ and supplement with 1 mL of 50 mg/mL carbenicillin. Mix well and pour ~ 25 mL per 10 cm petri plate. After 2 days incubation at room temperature, store plate inverted at 4°C packaged in original sleeves.
11. LB agar plates with carbenicillin: in a final volume of 1 L, dissolve 10 g of bacto tryptone, 5 g of yeast extract, 10 g of sodium chloride; once dissolved add 20 g of bacto agar and autoclave immediately. Cool to $\sim 50^{\circ}\text{C}$ before supplementing with 1 mL of 75 mg/mL carbenicillin. Mix well and pour ~ 25 mL per 10 cm petri plate. After 2 days incubation at room temperature, store inverted at 4°C packaged in original sleeves.
12. Plasmid DNA purification kit.

2.3 PCR-Mediated VEGAS Assembly

1. High-fidelity DNA polymerase and buffer.
2. dNTPs for PCR.
3. VEGAS acceptor vector: pJC170.
4. Agarose powder.
5. Agarose gel electrophoresis running buffer (e.g., TAE, TBE, TTE): dilute to $1\times$ concentration using deionized water. This solution does not need to be autoclaved.
6. $6\times$ DNA loading dye for gel electrophoresis: prepare a solution of 30% (v/v) glycerol and 0.25% (w/v) bromophenol blue in water. This solution does not need to be autoclaved.
7. DNA gel extraction kit.

8. Yeast strain: obtain a stock of the common lab yeast strain BY4741 [8] (*see Note 6*).
9. YPD liquid medium: dissolve 10 g of yeast extract and 20 g of bacto peptone in a final volume of 800 mL of water. Dissolve 20 g of dextrose in a final volume of 200 mL of water. After autoclaving both solutions individually, mix together for a final volume of 1 L.
10. Spectrophotometer and cuvettes for measuring cell density (OD).
11. 1 M lithium acetate: dissolve 65.98 g of lithium acetate in water and bring to a final volume of 1 L.
12. 0.1 M lithium acetate: dilute 100 mL of 1 M lithium acetate into 900 mL of water.
13. 44% PEG 3350 (w/v): dissolve 44 g of PEG 3350 in a final volume of 100 mL and filter-sterilize.
14. Carrier DNA: prepare 500 μ L aliquots of herring sperm DNA (10 mg/mL) and store at -20°C .
15. 1 M calcium chloride dihydrate: dissolve 14.7 g of calcium chloride dihydrate in a final volume of 100 mL of water.
16. 5 mM calcium chloride dihydrate: dilute 0.5 mL 1 M calcium chloride dihydrate in 95.5 mL of water.
17. Dimethyl sulfoxide (DMSO), molecular biology grade.
18. SC-Ura agar plates: In a final volume of 700 mL of water, dissolve 1.7 g yeast nitrogen base without amino acids and without carbohydrate and without ammonium sulfate, and 5 g ammonium sulfate; once dissolved add 20 g bacto agar. In a final volume of 200 mL of water, dissolve 20 g of dextrose. In a final volume of 100 mL, dissolve 2 g of synthetic drop-out powder without uracil. Autoclave the three solutions separately, mix together, and cool. Pour \sim 25 mL of medium per 10 cm petri plate. One liter of medium is enough for \sim 40 plates.
19. SC-Ura liquid medium: In a final volume of 800 mL of water, dissolve 1.7 g yeast nitrogen base without amino acids and without carbohydrate and without ammonium sulfate, 5 g ammonium sulfate, and 2 g of synthetic drop-out powder without uracil. In a final volume of 200 mL of water, dissolve 20 g of dextrose. Autoclave the two solutions separately, then mix together.

2.4 Plasmid Recovery from Yeast

1. Acid-washed glass beads, 0.5 mm.
2. P1 resuspension solution: 50 mM Tris-HCl pH 8.0, 10 mM EDTA, 100 μ g/mL RNase. Store at 4°C (*see Note 7*).

3. P2 lysis solution: 200 mM NaOH, 1% SDS (w/v). Do not autoclave and store at room temperature (*see Note 8*).
4. N3 neutralization solution: 4.2 M guanidinium hydrochloride, 0.9 M potassium acetate, pH 4.8 (*see Note 9*).
5. Benchtop shaker or vortex with adapter for microcentrifuge tubes.
6. Miniprep columns, recovery tubes, wash buffers, and elution buffer: these items can be appropriated from a DNA plasmid purification kit.

3 Methods

3.1 *yGG to Assemble Violacein Pathway TUs Compatible with VEGAS*

1. Design yGG assemblies to generate VEGAS-compatible transcription units. Each yGG assembly should include a promoter (PRO) and terminator (TER) to regulate expression of the coding sequence (CDS). Additionally, to enable PCR-mediated VEGAS, unique LVA and RVA parts should be assigned to each PRO-CDS-TER set in the pathway (*see Note 10*). The list of parts for the assembly of the violacein pathway is given in Table 2 (*see Note 11*).
2. Culture *E. coli* strains containing individual violacein pathway parts (Table 2) in 5 mL LB liquid medium with kanamycin at 30 °C for 18–24 h in a drum rotator (~200 rpm).
3. Culture the *E. coli* strain carrying the yGG acceptor vector (pJC120) in 5 mL LB liquid medium with carbenicillin at 30 °C for 18–24 h in a drum rotator (~200 rpm).
4. Prepare plasmid DNA from each culture using a plasmid DNA purification kit and quantify plasmid concentrations with a spectrophotometer. Calculate the volume of each part needed to provide the molar equivalent to 100 ng of the acceptor vector.
5. For each TU assembly, combine the following components in a PCR tube: 100 ng acceptor vector, equal moles of all other assembly parts, 1.5 µL 1 mg/mL BSA, 1.5 µL 10× T4 DNA ligase reaction buffer, 0.4 µL *Bsa*I, 1 µL T4 DNA ligase, and water to a final volume of 15 µL (*see Note 12*).
6. Mix the reactions by tapping the tubes and briefly centrifuge to collect samples at the bottom of the tubes.
7. Transfer tubes to a thermal cycler and incubate with the following program: 25 cycles of [3 min at 37 °C and 4 min at 16 °C], 1 cycle of [5 min at 50 °C and 5 min at 80 °C], and a hold at 4 °C (*see Note 13*).
8. Transform each assembly reaction into *E. coli*. In brief, thaw 50 µL competent cell aliquots on ice. Add 2 µL of assembly

reactions to thawed cells, tap to mix, and incubate on ice for 30 min. Heat shock samples for 45 s at 42 °C. Return cells to ice immediately and incubate for 2 min. Add 450 µL of LB medium and incubate for 30 min to 1 h at 30 °C. Plate the transformation products on LB agar plates with carbenicillin.

9. After incubation at 30 °C for 24–36 h (*see Note 14*), visually inspect the colonies. Those containing putative correct assemblies will appear white in color, compared to those containing uncut acceptor vector, which will appear pink/red to the naked eye due to expression of a red fluorescent protein in *E. coli* (*see Note 15*).
10. Inoculate 2–3 white colonies from each assembly into 5 mL of LB media with carbenicillin. Incubate cultures for 18–24 h at 30 °C in a drum rotator (~200 rpm).
11. Prepare plasmid DNA from each overnight culture using a plasmid DNA purification kit. Evaluate the assembly structure of each purified TU plasmid by restriction digestion using *BsmBI* and/or other restriction endonucleases, followed by agarose gel electrophoresis. Compare the *in silico* predicted digestion pattern to the observed pattern of band migration on the gel to identify correctly assembled TUs (Fig. 2).

3.2 Preparation of PCR Amplicons for PCR-Mediated VEGAS

1. Quantify the DNA concentration of the assembled TUs and dilute each in water to a final concentration of ~20 pg/µL.
2. Using a high-fidelity DNA polymerase in a 50 µL reaction volume, amplify each TU with primers designed to introduce terminal homology between adjacent TUs. A list of primers for PCR-mediated VEGAS violacein pathway assembly is provided in Table 3.
3. Mix the reactions by tapping the tubes and briefly centrifuge to collect samples.
4. Transfer tubes to a thermal cycler and incubate with the following program: 1 min at 98 °C, 30 cycles of [15 s at 98 °C, 30 s at 55 °C, 2 min 30 s at 72 °C], a final extension of 5 min at 72 °C, and a hold at 4 °C (*see Note 16*).
5. Linearize the VEGAS acceptor vector (pJC170). In a microcentrifuge tube or PCR tube, combine the following components: 22 µL pJC170 miniprep DNA, 3 µL 10× NEB Cutsmart buffer, 2 µL *BsaI*, 3 µL water. Incubate at 37 °C for 1 h.
6. Using the entire PCR product and the entire acceptor vector digest, add DNA loading dye to a final concentration of 1× and perform gel electrophoresis using a 1% agarose gel. Verify that each amplified TU is the expected size and that the acceptor vector digestion produces two bands at 5 kb and 950 bp. Excise each amplified TU using a clean razor blade and transfer the

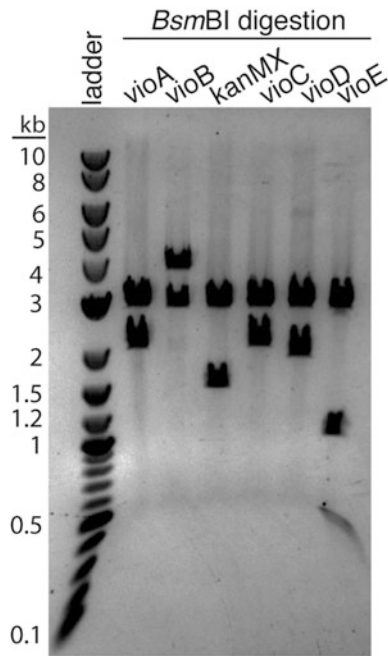


Fig. 2 Digestion verification of yGG assembled TUs. White colonies derived from each of six violacein yGG assembly reactions (vioA (pJC094), vioB (pJC100), KanMX (pJC070), vioC (101), vioD (pJC095), vioE (pJC097)) were grown in culture overnight and TU plasmids prepped. Migration of *BsmBI* digestion products separated by 1% agarose gel electrophoresis is shown. Ladder bands (in kb) are indicated on the *left*. *BsmBI* digestion of all assembled TUs yields a vector band ~3 kb and the intact, assembled TU. The single exception, vioE, yields two bands in addition to the vector band due to the spurious creation of a *BsmBI* site at the CDS-TER assembly junction

Table 3
Primers for PCR-mediated VEGAS violacein assembly

| TU | For primer sequence | Rev primer sequence |
|-------|---|---|
| vioA | GAT CTG TCC TTC TCT GCC GGC GAT CGT AGT TAT TAG CGG CGC CGG GAA ATC CAG CAT ATT | ATG CGA GTT CAG GTG TAA CGT AGA CTC ACA GTG GCC GCG AGT CAC TAC TAA ACT GTC TCT |
| vioB | AGA GAC AGT TTA GTA GTG ACT CGC GGC CAC TGT GAG TCT ACG TTA CAC CTG AAC TCG CAT | CAC ATC AGG CCC GGG CGG CCA GTA CGC CCG CCC TGG GGA CAC CAT CCC GCC TCA ATC ACA |
| kanMX | TGT GAT TGA GGC GGG ATG GTG TCC CCA GGG CGG GCG TAC TGG CCG CCC GGG CCT GAT GTG | AGT ACC GGA TCC TAA AGC CGA TTC AAG GGA CTT TCT ATC AGC CCG ACG ACA TGC ATC GTC |
| vioC | GAC GAT GCA TGT CGT CGG GCT GAT AGA AAG TCC CTT GAA TCG GCT TTA GGA TCC GGT ACT | GAG AGT TTA CAC CTC TTC AAA ACT TGC CGA AAT AGC CGC AAT CGT ATC TGA CTT CTC CCA |
| vioD | TGG GAG AAG TCA GAT ACG ATT GCG GCT ATT TCG GCA AGT TTT GAA GAG GTG TAA ACT CTC | GGC TTG TCG ATG CAC GCA CCC GCG ATA AGC AGT ACA TTG GCG CGG TCT ACG TCT ATG GAA |
| vioE | TTC CAT AGA CGT AGA CCG CGC CAA TGT ACT GCT TAT CGC GGG TGC GTG CAT CGA CAA GCC | CGT ACG GGG TCA CGC ATC CAA GCG TCA TAG AAA ACG CGC TCA AAG CGA TGA CGC GGA CTA |

excised gel slice to a microcentrifuge tube. Excise the 5 kb band of the linearized VEGAS acceptor vector digest and transfer the gel slice to a microcentrifuge tube.

7. Perform gel purification of the TU amplicons and linearized VEGAS acceptor vector according to the manufacturer's instructions.
8. Quantify DNA concentration of gel purified VEGAS acceptor vector and amplified TUs with a spectrophotometer.

3.3 VEGAS Assembly of Violacein Pathway in Yeast

1. Inoculate a single colony of the yeast strain to be used for VEGAS assembly into 5 mL of YPD medium. Incubate overnight at 30 °C with agitation (*see Note 17*).
2. Measure the OD₆₀₀ of the overnight yeast culture in a spectrophotometer. From the overnight culture, inoculate a new culture at an OD₆₀₀ of 0.1. For each transformation reaction to be performed prepare 5 mL of sub-cultured cells (e.g., for four transformations, subculture in a volume of 20 mL). Incubate for 4–5 h at 30 °C in a drum rotator (~200 rpm). Centrifuge the yeast culture (4500 × *g*, 3 min). Discard the liquid and resuspend the cells in an equal volume of sterile water.
3. Centrifuge to pellet the cells (4500 × *g*, 3 min). Discard the liquid and resuspend the cells in an equal volume of 0.1 M lithium acetate.
4. Centrifuge to pellet the cells (4500 × *g*, 3 min). Discard the liquid and transfer yeast cells to microcentrifuge tube in the remaining supernatant by pipetting.
5. Centrifuge to pellet the cells (4500 × *g*, 3 min). Remove supernatant by aspirating or pipetting.
6. Resuspend cells in 0.1 M lithium acetate. The total volume should be 40 µL per transformation.
7. Prior to preparing the yeast transformation mixture in the next step, boil an aliquot of herring sperm DNA at 95 °C for 5 min and subsequently incubate it on ice for at least 5 min.
8. Prepare the yeast transformation mixture. For a single transformation reaction combine 240 µL 44% PEG-3350, 36 µL 1 M lithium acetate, 25 µL 10 mg/mL herring sperm DNA (*see step 7*), and 40 µL yeast cells in 0.1 M lithium acetate in a microcentrifuge tube and vortex for 10 s. For multiple transformation reactions, prepare the transformation as a master mix and aliquot 340 µL of the transformation mix to the appropriate number of microcentrifuge tubes.
9. Add 50 ng gel-purified *Bsa*I-linearized VEGAS acceptor vector, and ~200 ng of each gel-purified TU PCR product. Vortex tubes for 10 s and incubate the tubes for 30 min at 30 °C (*see Note 18*).

10. Add 36 μL of DMSO to each transformation tube and vortex to mix.
11. Heat shock the samples at 42 °C for 15 min.
12. Centrifuge samples for 3 min at 4500 $\times g$ and aspirate the supernatant leaving the cell pellet intact.
13. Resuspend cells in 400 μL of 5 mM calcium chloride by pipetting up and down.
14. Plate 350 μL and 50 μL of the transformation on separate SC-Ura plates.
15. Incubate plates at 30 °C for 3–5 days, until colonies are visible. Yeast colonies carrying correctly assembled violacein pathways should develop a purple/black color.
16. Re-streak 4–8 single transformants on fresh SC-Ura plates for each transformation to single colony purify.

3.4 Recovery of Assembled Pathways from Yeast into *E. coli*

1. For each assembly to be recovered, inoculate a streak-purified single yeast colony into 3–5 mL of SC-Ura liquid medium. Incubate overnight at 30 °C with rotation.
2. Transfer 1.5 mL of overnight culture to a microcentrifuge tube. Spin cells down in a microcentrifuge, 3 min at 4500 $\times g$. Aspirate the supernatant.
3. Resuspend the pelleted cells in 250 μL P1 resuspension buffer.
4. To the cell suspension, add 0.5 mm acid-washed glass beads until the beads reach approximately the 200 μL mark of the microcentrifuge tube.
5. In a benchtop shaker or vortex with an adapter for microcentrifuge tubes, shake the tubes for 10 min at room temperature.
6. Add 250 μL of P2 lysis buffer. Invert the tubes three times to mix and incubate for 5 min at room temperature.
7. Add 350 μL of N3 neutralization buffer. Invert the tubes ten times to mix.
8. Centrifuge 10 min at 11,000–16,000 $\times g$.
9. Transfer the supernatant to a miniprep column placed in a collection tube. Centrifuge for 1 min, discard the flow-through, and place the column back into the collection tube.
10. Wash the miniprep column according the manufacturer's instructions for *E. coli* minipreps.
11. Transfer the column to a microcentrifuge tube. Add 30 μL of elution buffer from the plasmid DNA purification kit to the column and let stand for 1 min.
12. Centrifuge for 1 minute at 11,000–16,000 $\times g$ to elute the DNA from the column.

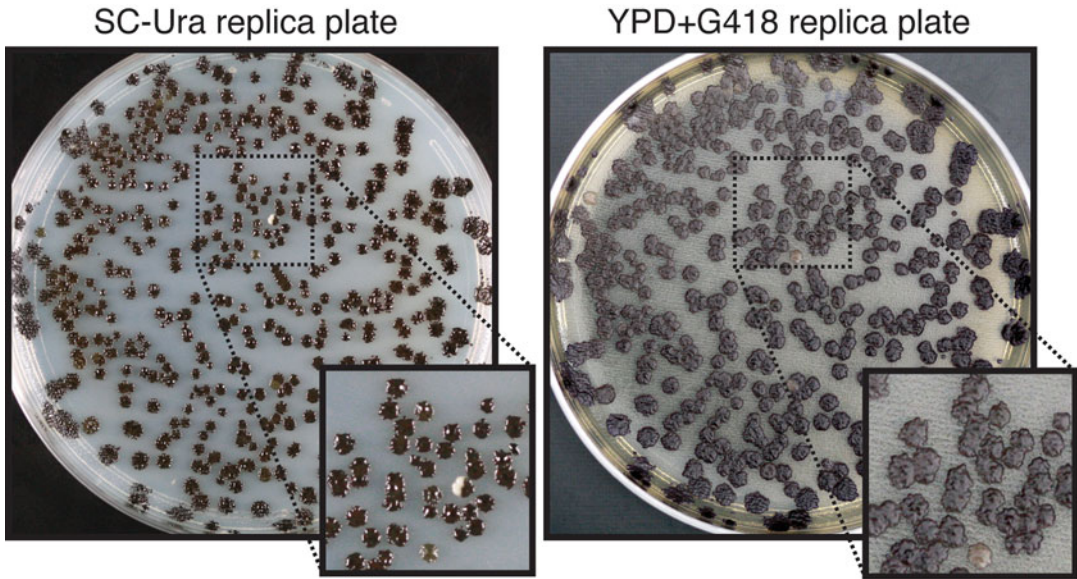


Fig. 3 Violacein pathway expression in *S. cerevisiae*. The yGG/VEGAS assembled violacein pathway was first recovered into *E. coli* and digestion verified before retransforming *S. cerevisiae*. Replica plating of the primary transformants onto SC-Ura and YPD+G418 is shown to demonstrate the genetic stability of the pathway based on overall uniformity of color

13. Transform the eluted DNA product into chemically competent *E. coli* cells as in **step 7** of Subheading 3.1 except using 10 μL of eluted DNA and 100 μL of cells. The outgrowth step following heat shock should use 900 μL of LB liquid medium and the entire transformation product should be plated equally on 2 LB agar plates with carbenicillin.
14. Inoculate 4 individual *E. coli* transformants into 5 mL LB liquid medium with carbenicillin and incubate in a drum rotator (~200 rpm) for 18–24 h at 30 °C.
15. Prepare plasmid DNA from each overnight culture using a plasmid DNA purification kit. Evaluate the assembly structure of each purified TU plasmid by restriction digestion, followed by agarose gel electrophoresis. Compare the in silico digestion pattern to the observed pattern of band migration on the gel to identify correctly assembled pathways.
16. Correctly assembled constructs can be retransformed into yeast; nearly all colonies should appear dark purple/black in color (Fig. 3).

4 Notes

1. VEGAS-compatible transcription unit parts include left and right VEGAS adapters (LVA and RVA, respectively), yeast promoters (PRO), coding sequences (CDS), and yeast

terminators (TER). Parts are flanked by inwardly facing type IIS restriction endonuclease sites (e.g., *BsaI*, *BsmBI*) that leave the 5' and 3' overhangs listed in Table 1 upon cleavage. Recognition sites for the type IIS restriction enzyme used for Golden Gate assembly should otherwise be absent within each part. Parts can be cloned into vectors with a drug resistance that is different from the yGG acceptor vector (pJC120 encodes ampicillin resistance) or may be PCR products. To remove undesired type IIS restriction enzyme recognition sites from parts or vectors, we use a previously described method called "MISO" [9].

2. 10× T4 DNA ligase reaction buffer can be purchased from NEB (B0202S).
3. To maximize the units of *BsaI* enzyme introduced into a yGG reaction and simultaneously minimize the amount of glycerol, we recommend using a high concentration version of *BsaI*, for example *BsaI*-HF (NEB, R3535), which is twice as concentrated as the standard stock (NEB, R0535). Alternatively, NEB will prepare a tenfold concentrated stock by special request.
4. *E. coli* competent cells used for transforming yGG reaction products should have a transformation efficiency of $>10^6$ colony forming units per μg of DNA transformed. These can be purchased or made in the lab [10].
5. Carbenicillin is a semisynthetic ampicillin analog and is used as a substitute to confer resistance to the β -lactamase gene as it tends to be more stable.
6. Any *S. cerevisiae* strain competent for homologous recombination is appropriate to use for VEGAS pathway assembly and expression. If the desired *S. cerevisiae* strain is prototrophic for all biosynthetic pathways, a drug resistance marker can be encoded on the assembly vector for selection. BY4741 can be purchased (GE Dharmacon, YSC1048) or obtained by request from yeast labs around the world.
7. The P1 resuspension buffer can be purchased (Qiagen, 19051). The RNase is sold separately (Qiagen, 19101).
8. The P2 lysis buffer can be purchased (Qiagen, 19052).
9. The N3 neutralization buffer can be purchased (Qiagen, 19064). We have found it important to use N3 as opposed to P3 to ensure recovery of *E. coli* transformants.
10. In this protocol we describe PCR-mediated VEGAS, which enables specification of TU order and orientation after TUs are assembled by yGG. This is achieved by assigning unique LVA and RVA parts to each TU (*see* Table 2); the generation of terminal homology between adjacent parts is generated by PCR amplifying assembled TUs using primers with overhangs

that encode terminal homology. An alternate approach for assembly is to specify terminal homology between adjacent parts in the assignment of LVA and RVA parts. This strategy, termed adapter-mediated VEGAS, is described in a previous publication, and requires the TU order and orientation be determined prior to yGG TU assembly [7].

11. It is often useful to encode a yeast selectable marker TU in addition to the set of TUs required for expression of the pathway of interest. This enables facile screening of VEGAS-derived yeast transformants by replica plating onto the appropriate selective medium. In the case of violacein assembly described here, we encode a kanMX TU as the third gene in the pathway (Table 2). The kanMX donor part for yGG already has built in PRO and TER regulatory sequences, so these parts are left out of this yGG reaction (Table 2).
12. It is most efficient to prepare a master mix of all common reagents including the acceptor vector, which can be aliquoted to PCR tubes, and subsequently add the TU donor parts as defined.
13. A faster, alternative protocol for yGG assembly may be carried out as follows: 60 min at 37 °C, 5 min at 50 °C, 10 min at 80 °C, and a hold at 4 °C. The number of red colonies post-yGG assembly is usually higher using this protocol.
14. *E. coli* transformation plates may also be incubated at 37 °C, which promotes faster colony growth and increased red color development. However, if the assembled TU is not well tolerated by *E. coli*, incubation at 30 °C can minimize potential rearrangements/loss of insert DNA.
15. *E. coli* colony color develops more slowly at 30 °C. If a fluorescent imager is available this can enable red/white colony screening at earlier time points.
16. The extension time of the PCR reaction should be set with respect to the longest TU to be amplified; in the case of violacein this is vioB (Table 2).
17. The yeast transformation protocol described here is based on a previous publication [11] with minor modifications. Yeast competent cells can also be prepared in advance and stored at -80 °C [12].
18. It is advantageous to include several control transformations. As we say in the Boeke lab—“No control? Out of control.” Always include a positive control (20 ng of e.g., circular pJC170) to calculate transformation efficiency for the experiment (colony forming units/μg DNA transformed) and a negative control (no DNA) where no transformants are expected. We typically include a “linear vector alone” control to assess

the number of colonies that may be arise due to vector reclosure by non-homologous end joining. We also include a control reaction “missing one part”, where a single TU is intentionally left out of the transformation reaction; this can be helpful to troubleshoot mis-assemblies.

Acknowledgements

This work was supported by the National Science Foundation grant MCB-0718846 and Defense Advanced Research Projects Agency grant N66001-12-C-4020.

References

1. Verwaal R, Wang J, Meijnen JP, Visser H, Sandmann G, van den Berg JA, van Ooyen AJ (2007) High-level production of beta-carotene in *Saccharomyces cerevisiae* by successive transformation with carotenogenic genes from *Xanthophyllomyces dendrorhous*. *Appl Environ Microbiol* 73(13):4342–4350. doi:[10.1128/AEM.02759-06](https://doi.org/10.1128/AEM.02759-06)
2. Choi JW, Da Silva NA (2014) Improving polyketide and fatty acid synthesis by engineering of the yeast acetyl-CoA carboxylase. *J Biotechnol* 187:56–59. doi:[10.1016/j.jbiotec.2014.07.430](https://doi.org/10.1016/j.jbiotec.2014.07.430)
3. Koopman F, Beekwilder J, Crimi B, van Houwelingen A, Hall RD, Bosch D, van Maris AJ, Pronk JT, Daran JM (2012) De novo production of the flavonoid naringenin in engineered *Saccharomyces cerevisiae*. *Microb Cell Fact* 11:155. doi:[10.1186/1475-2859-11-155](https://doi.org/10.1186/1475-2859-11-155)
4. Krivoruchko A, Nielsen J (2014) Production of natural products through metabolic engineering of *Saccharomyces cerevisiae*. *Curr Opin Biotechnol* 35C:7–15. doi:[10.1016/j.copbio.2014.12.004](https://doi.org/10.1016/j.copbio.2014.12.004)
5. Boone C (2014) Yeast systems biology: our best shot at modeling a cell. *Genetics* 198(2):435–437. doi:[10.1534/genetics.114.169128](https://doi.org/10.1534/genetics.114.169128)
6. Agmon N, Mitchell LA, Cai Y, Ikushima S, Chuang J, Zheng A, Choi WJ, Martin JA, Caravelli K, Stracquadanio G, Boeke JD (2015) Yeast golden gate (yGG) for the efficient assembly of *S. cerevisiae* transcription units. *ACS Synth Biol* 4(7):853–859. doi:[10.1021/sb500372z](https://doi.org/10.1021/sb500372z)
7. Mitchell LA, Chuang J, Agmon N, Khunsriraksakul C, Phillips NA, Cai Y, Truong DM, Veerakumar A, Wang Y, Mayorga M, Blomquist P, Satta P, Trueheart J, Boeke JD (2015) Versatile Genetic Assembly System (VEGAS) to assemble pathways for expression in *S. cerevisiae*. *Nucleic Acids Res* 43(13):6620–6630. doi:[10.1093/nar/gkv466](https://doi.org/10.1093/nar/gkv466)
8. Brachmann CB, Davies A, Cost GJ, Caputo E, Li J, Hieter P, Boeke JD (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(2):115–132. doi:[10.1002/\(SICI\)1097-0061\(19980130\)14:2<115::AID-YEA204>3.0.CO;2-2](https://doi.org/10.1002/(SICI)1097-0061(19980130)14:2<115::AID-YEA204>3.0.CO;2-2)
9. Mitchell LA, Cai Y, Taylor M, Noronha AM, Chuang J, Dai L, Boeke JD (2013) Multi-change isothermal mutagenesis: a new strategy for multiple site-directed mutations in plasmid DNA. *ACS Synth Biol* 2(8):473–477. doi:[10.1021/sb300131w](https://doi.org/10.1021/sb300131w)
10. Inoue H, Nojima H, Okayama H (1990) High efficiency transformation of *Escherichia coli* with plasmids. *Gene* 96(1):23–28
11. Gietz RD (2014) Yeast transformation by the LiAc/SS carrier DNA/PEG method. *Methods Mol Biol* 1205:1–12. doi:[10.1007/978-1-4939-1363-3_1](https://doi.org/10.1007/978-1-4939-1363-3_1)
12. Gietz RD, Schiestl RH (2007) Frozen competent yeast cells that can be transformed with high efficiency using the LiAc/SS carrier DNA/PEG method. *Nat Protoc* 2(1):1–4. doi:[10.1038/nprot.2007.17](https://doi.org/10.1038/nprot.2007.17)

Part III

Screening and Selection of Synthetic Metabolic Pathways

Multi-capillary Column Ion Mobility Spectrometry of Volatile Metabolites for Phenotyping of Microorganisms

Christoph Halbfeld, Jörg Ingo Baumbach, Lars M. Blank, and Birgitta E. Ebert

Abstract

Rational strain engineering requires solid testing of phenotypes including productivity and ideally contributes thereby directly to our understanding of the genotype–phenotype relationship. Actually, the test step of the strain engineering cycle becomes the limiting step, as ever advancing tools for generating genetic diversity exist. Here, we briefly define the challenge one faces in quantifying phenotypes and summarize existing analytical techniques that partially overcome this challenge. We argue that the evolution of volatile metabolites can be used as proxy for cellular metabolism. In the simplest case, the product of interest is a volatile (e.g., from bulk alcohols to special fragrances) that is directly quantified over time. But also nonvolatile products (e.g., from bulk long-chain fatty acids to natural products) require major flux rerouting that result potentially in altered volatile production. While alternative techniques for volatile determination exist, rather few can be envisaged for medium to high-throughput analysis required for phenotype testing. Here, we contribute a detailed protocol for an ion mobility spectrometry (IMS) analysis that allows volatile metabolite quantification down to the ppb range. The sensitivity can be exploited for small-scale fermentation monitoring. The insights shared might contribute to a more frequent use of IMS in biotechnology, while the experimental aspects are of general use for researchers interested in volatile monitoring.

Key words Ion mobility spectrometry, Multi-capillary column, Online analysis, On-site analysis, Phenotype screening, Volatile metabolites, Volatile organic compounds

1 Introduction

1.1 Challenge of the Test Step

The development of cell factories for the economically viable production of industrially interesting compounds has been significantly sped up in recent years. This has been achieved by the development of efficient workflows and automation of single-steps of the so-called Design–Build–Test–Learn (DBTL) cycle [1]. Especially, the design and build steps have radically been accelerated by the development of dedicated software tools, rapid genome modification methods, the enormous drop in gene synthesis costs and robotics systems that automate the single steps from

PCR amplification to parts assembly, transformation, and genome integration allow today to generate thousands of strains in one single week [2]. With these tools at hand it is now possible to not only access Nature's biochemical capabilities but also enhance the biochemical space by increasing the genetic diversity and ultimately exploit it for the generation of superior production strains. However, the subsequent steps of the engineering cycle—Test and Learn—are lacking behind and are impeding speedy and seamless rounds through the DBTL cycle and hence overall cut-down of strain optimization times but also rational engineering approaches that are guided by *in silico* simulations performed on the basis of the data generated in the “Test” step.

The lower capacity of the Test step is grounded in the low generalizability of analytical methods for the broad diversity of chemical compounds to be detected and quantified, requiring assay development on a case-by-case basis, the inherently limited throughput of standard analytical techniques, such as chromatography-based methods and the trade-off between assay throughput and sensitivity as well as information content [3]. For an initial phenotypic characterization of an engineered strain low information content is acceptable as strain selection criteria are based on growth behavior and production performance, which do not require systems-wide omics data. The latter are required for in-depth analyses of selected overproducers to provide insight about potential bottlenecks that can be overcome in a subsequent DBTL cycle. Although important for the whole strain engineering process, we here do not touch upon omics techniques but limit the following overview to methods for phenotypic screening, with which important performance parameters such as yield per gram substrate or biomass, production rate, and final titer can be determined. Accordingly, the techniques described below focus on measurements of biomass, carbon substrate, and the target molecule.

1.2 Existing Techniques

As mentioned above phenotypic screens cannot easily be generalized but are rather one-offs specialized for the specific analyte and parameter to be determined, e.g., endpoint titers vs. kinetic profile of product formation, with the latter being preferred as it allows deeper insight into the physiology and catalytic performance of the host.

Ideally the target molecule is a natural chromophore or fluorophore, hence can directly be measured in plate-based assays, which allow online analysis with great throughput and often with high sensitivity. However, most target molecules do not confer a clear phenotype, exceptions being colorful carotenoids such as lycopene and astaxanthin [4, 5]. Inconspicuous molecules that emit no signal either have to be converted into detectable outputs or alternative measurement techniques have to be employed. We here summarize often used methods and discuss them with respect to throughput, sensitivity, and dynamic range.

1.2.1 *Classic Techniques (HPLC, GC)*

Since decades, fermentation broth is analyzed by liquid and/or gas chromatography (GC) coupled to appropriate detectors. A clear preference for high-performance liquid chromatography (HPLC) coupled to UV/RI (ultraviolet/refractive index) detectors exist in the literature, as GC analysis requires most often sample preparation to remove water and increase volatility of the analytes. An exception is the use of a multi-capillary column (MCC) as explained in more detail in the ion mobility spectrometry section and the protocol. For fermentation broth analysis by HPLC, a clear preference for the Aminex HPX-87H column from Bio-Rad Laboratories (Hercules, CA, USA) exists, as this column has a high resolution for sugars, alcohols, and carboxylic acids (although cheaper, but most often worse performing alternatives exist). While the sugars and alcohols are detected using the RI detector, the carboxylic acids are quantified by either the UV or if the concentration is high enough, the RI detector. Most RI detectors are comparably insensitive, with limits of quantification in the high micromolar or low to medium millimolar range, depending on the analyte of interest. The performance of UV detectors depends strongly on the absorption coefficient of the analyte. Dicarboxylic acids like fumarate can be quantified in the low micromolar range. The sample throughput is limited, as the column is rather long and fragile; allowing only limited pressure, resulting easily in single run times of 20 and more minutes. Combined with an autosampler, still up to 72 samples can be analyzed per day. Hence, for many microbes and growth media used, the main substrates and products can be quantified by a single HPLC-UV-RI method.

For GC, flame ionization detectors (FIDs) are the most often used, in which the analyte is chemically ionized and the formed ions are detected. Most often, less hydrophilic metabolites are extracted from the growth medium using an organic solvent that is subsequently dried. Alternatively, the entire growth medium is dried. Depending on the volatility, the analytes can be directly injected onto the GC column, like small- and medium-sized fatty alcohols, or have to be derivatized before GC-FID analysis, like medium- and long-chain fatty acids. The separation power is high as peak broadening is often low on GC columns. In GC-FID analysis, the limit of quantification depends on the absolute number of carbon atoms ionized as this is proportional to the signal measured and can be in the low micromolar range or lower. The runtime depends on many parameters including the analyte and the length of the column, it varies on generally used systems between 10 and 30 min. Since autosamplers are commonly used, medium-throughput analyses are possible with GC-FID systems.

These traditional analytical techniques allow the measurement and quantification of many of the substrates and products encountered in microbiology. However, throughput, online or in situ measurements, and the possibility of analyte identification is

somewhat cumbersome or just not possible. Hence, we briefly summarize some additional techniques that cover some of the challenges encountered when characterizing the phenotype of microbes.

1.2.2 Chemical or Enzymatic Transformation Assays

Molecules not easily, directly measurable, might be converted to a detectable compound by reaction with an exogenously added chemical or by enzyme catalysis. Alternatively, a stoichiometric byproduct (e.g., H_2O_2), or the conversion of a cofactor (ATP, NAD(P)H, CoA) can be quantified. Reaction-associated pH changes can also be linked to the concentration of the compound of interest. A prerequisite for these kinds of assays is that the coupling reactions are irreversible and are driven to completion to allow accurate back-calculation of the product concentration. Known examples are commercial enzyme assay kits for glucose, acetate or succinate, with colorimetric or fluorescent readout, which excel in high sensitivities in the nmol/well to pmol/well and low sample volume requirements (1–50 μ L).

These transformations can easily be parallelized in microtiter plate format and with novel acoustic liquid dispenser like the Echo liquid handler from Labcyte (Sunnyvale, CA; USA), can be scaled-down to 384- or even 1536-well format given the assay for the end product or the converted cofactor is sensitive enough. TRACE Analytics GmbH, Braunschweig, Germany, provides a device for online bioreactor monitoring of glucose, lactate, methanol, and ethanol. Samples are either taken from the reactor using filtration probes or volume-free by dialysis probes or sample tubing. In the dialysis-based sampling device the fermentation broth is continuously pumped along/through a semipermeable membrane or tube; the analyte diffuses through the membrane until equilibrium is established. The analytes are then converted by a membrane-immobilized oxidase and the reduction of the oxygen concentration is measured by an amperemeter.

Although throughput is comparably high, the requirement of an isolated enzyme with limited life per analyte investigated, stoichiometric amounts of cofactors or chemicals render these assays expensive and hence often impractical for the screening of large libraries.

1.2.3 Advanced Spectroscopy Including Raman, Near Infrared, and Nuclear Magnetic Resonance

Ideally, one would like to determine substrates, products, and biomass at real time, noninvasively in any sized microbial culture. Some spectroscopy techniques including Raman, near infrared, and nuclear magnetic resonance spectroscopy have the potential to cover some of the aspects one is wishing for. Raman spectroscopy was for example used to determine carotenoid synthesis by recombinant *Saccharomyces cerevisiae* noninvasively at almost real time [6]. The total lipid content correlated in this study with

carotenoid synthesis, and both were quantifiable via Raman spectroscopy. In a different example, itaconate synthesis from glucose by the smut fungus *Ustilago maydis* was quantified by low-field NMR spectroscopy [7]. The fermenter was equipped with a measuring loop connected to the low-field NMR. The setup allowed noninvasive quantification of glucose uptake and itaconate synthesis in a complex growth medium that contained yeast extract, besides the obvious biomass of the fungus. For strain phenotyping or process optimization, the number of experiments should be high. Just recently, a workflow for noninvasive fermentation monitoring using near-infrared spectroscopy was reported [8]. The workflow was demonstrated using another fungus, the challenging filamentously growing *Penicillium chrysogenum*. The root mean square errors for biomass, penicillin, phenoxyacetic acid, and ammonia were about 2.6 g/L, 0.3 g/L, 0.5 g/L, and 18 mM, respectively. The benefit of this noninvasive measurement is immediately obvious for everyone, who monitored fermentation kinetics using individual assays for the very different analytes of interest.

While spectroscopy techniques will advance, their complicated use including data interpretation by experts, the often low resolution, sensitivity, and time required until the signals from a fermentation broth are deciphered in a cycle of experimental testing and data interpretation, will limit their use in phenotyping of microbes to dedicated applications, in which the number of analytes is low and ideally nonchanging.

1.2.4 Metabolite Sensors (GFP Based)

Fluorescent proteins are widely used for high-throughput characterization of parts, promoters or ribosome binding sites [9–12]. Fluorescent proteins with nonoverlapping extinction/emission wavelengths exist, green, red, yellow fluorescent proteins (GFP, RFP, YFP) allowing the co-expression of multiple reporters.

Fluorescent proteins can also be used as biosensors by exploiting the plethora of natural sensor devices, e.g., transcription factors, riboswitches, or enzymes that detect specific metabolites. For example, transcription factor-based biosensors have been used to control reporter gene expression in response to an effector molecule, here the target product. In this way the fluorescence intensity of the expressed reporter protein in a cell population or in single cells is correlated with the intracellular concentration of the molecule of interest. Transcription factor-based biosensors have successfully been used to monitor intracellular concentrations of amino acids [13], butanol [14] and flavonoids [15]. The fluorescence output can further be exploited for the selection of improved production strains using fluorescence-activated cell sorting (FACS), which was shown for example to isolate branched-chain amino acids overproducing *Corynebacterium glutamicum* from a mutant library generated by chemical mutagenesis [16].

RNA aptamer-based biosensors are a fusion of two aptamers. One of these aptamers binds a fluorophore (output domain) while the second specifically binds the metabolite of interest (sensing domain). Binding of the metabolite results in a conformational change of the aptamer, which alters the fluorescence of the fluorophore. As the response is immediate, RNA aptamer-based sensors are specifically useful to monitor dynamic changes in intracellular metabolite concentrations [17].

While very powerful, the application of these biosensors is today restricted due to the limited availability of known transcription factors or aptamers that specifically bind the metabolite of interest.

1.2.5 Measuring Volatiles Produced by Microbes

Volatiles are an interesting group of molecules that have the ability to evaporate at room temperature. Although this feature is the basis of the nomenclature, the molecular structure is diverse, covering acids, alcohols, aldehydes, aromatics, ketones, terpenes, and others. Hence, volatiles include many molecules with industrial interest. Ethanol is a prime example. The challenges for quantification are the high volatility from the growth medium and the often low concentrations observed. In online monitoring of volatile metabolites, the transfer line into the analytical equipment should be heated, to prevent condensation of metabolites with medium to high boiling points. While much research is published by food researchers (e.g., wine, beer), volatiles are rarely investigated in microbial biotechnology. As the amount in the off-gas corresponds to the concentration in the growth medium, the analysis of these molecules can be performed noninvasively.

Here, we briefly summarize some of the techniques used (for a recent overview of analytical techniques and the volatiles of yeast please refer to [18]), before introducing ion mobility spectrometry in greater detail, as this is the technique on which the presented protocol is based.

GC-MS—Gas Chromatography-Mass Spectrometry

Besides GC-FID, GC can be coupled to a mass detector that allows analyte identification by mass or if fragmentation is possible, by the mass distribution of the fragments. Several mass spectrometry (MS) detector principles exist that differ in price, mass resolution, and analysis time, among many other parameters. Examples are the most often used quadrupole and ion trap detectors, time of flight, and Orbitrap detectors. The latter two have high mass resolution possibilities. To take full advantage of the different techniques, hybrid MS, in which two or more MSs are coupled, exist.

For GC-MS analysis, as for all analytical techniques of volatiles, different sample preparations exist. The direct connection of a GC-MS to the headspace of a microbial culture is rather rare, as the setup is not simple. Still, this would be the preferred technique, as minimal sample handling is required.

A more frequently used technique is the extraction of volatile organic compounds (VOCs) from the growth medium or the off-gas of a bacterial culture. Here, several techniques exist covering solid and liquid phase extraction. Extraction onto a solid surface, the so-called solid phase extraction (SPE), is a commonly used technique. In SPE, the SPE material is exposed to the sample of interest, then covered to avoid desorption during transport to the GC-MS, and finally put into a thermal desorption device. Here, the SPE material is heated to desorb the VOCs that are consequently separated on the GC column and detected by the MS, hence the alternative name is thermal desorption spectroscopy (TDS). Solid phase microextraction, the miniaturization of SPE, uses a thin needle consisting of adsorbents. Further miniaturization can be achieved using direct immersion single drop microextraction (DI-SDME). Here, a drop of a water-immiscible extractant is exposed to the growth medium, and when the equilibrium of the analytes is achieved injected into the GC-MS. A variant of SDME for gas-liquid extraction is continuous flow microextraction (CFME), in which for example the off-gas of a microbial culture is passing the extractant. Again, the droplet is directly injected into the GC-MS for VOC analysis.

The described techniques have the advantage of a chromatographic separation, followed by a mass measurement, allowing the analysis of many analytes in a sample. Indeed, it was reported that two-dimensional GC-MS (GCxGC-MS) can be used for hundreds or even thousands of analytes if required [19]. The measurement of many VOCs in a sample, however, can be cumbersome, as analytes might be lost during sampling and transport. In the next two paragraphs techniques suited for online measurements are described.

SESI-MS—Secondary Electrospray Ionization- Mass Spectrometry

A rather powerful analytical technique for VOC determination is the use of a secondary electrospray ionization (SESI) source coupled to a mass analyzer [20], ideally a high mass resolution Orbitrap MS. The MS enables identification and quantification of analytes. Sample ionization via SESI proceeds in two steps. In the first step 0.1% formic acid in water is ionized by a regular electrospray to form an aerosol of charged, minute droplets. In the second step the analytes collide with these droplets and the secondary ionization occurs through ion–molecule interaction. The ionized analytes are guided into the MS. A very impressive real-time analysis by SESI-MS reported the day–night cycle of about 400 volatile metabolites of the plant *Begonia semperflorens* [21].

IMS—Ion Mobility Spectrometry

Ion mobility spectrometry (IMS) is for many frequent travelers a well-known analytical instrument, although less-known for the analysis principal, but rather from the sampling procedure, as it is

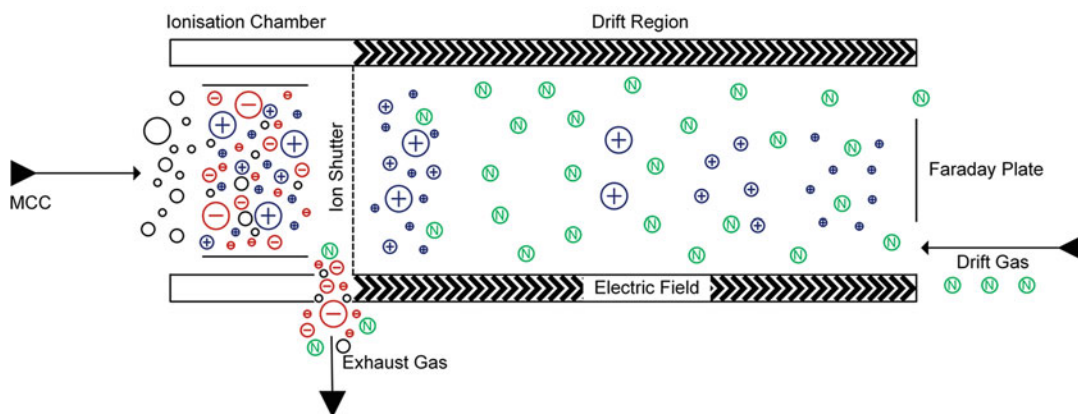


Fig. 1 Working principle of an ion mobility spectrometer. See text for explanation. *MCC* multi-capillary column. Modified from [23], previously published in [24]

used to analyze the molecules present on a swab of a trolley or bag at the airport's security control. Indeed, IMS has originally been used to detect explosives, chemical warfare agents or illegal drugs, and increasingly finds now applications in the life sciences for example in breath analysis [22] and in combination with MS detectors for peptide analysis. Performance features of IMS are the high sensitivity (detection limits down to pg/L to ng/L to or ppb_v/ppt_v) and high-speed data acquisition; a reading of a single spectrum takes only 20–50 ms. The cost of investment and maintenance is compared to most MS equipment low, because the IMS operates at ambient pressure and hence, no turbo vacuum pump is required. The IMS separates analytes by their gas-phase ion mobility. The analytes are ionized in the IMS, here, a drift tube IMS equipped with a ⁶³Ni radioactive ionization source (Fig. 1, see also protocol) is used, via charge transfer from ionized reactant ions. The ionized VOCs are guided by an electric field, while the entrance into the drift tube for analyte separation is separated from the ionization chamber by a shutter. The shutter opens periodically (a parameter that can be influenced for optimal analysis) to release the ionized analytes into the drift tube. In the drift tube, the analytes are accelerated by an electric field towards the detector, a Faraday plate. To increase separation, a drift gas (here highly pure nitrogen) flowing into the opposite direction is used (Fig. 1). The ion mobility of an analyte depends besides other molecule parameters on the mass, charge, and shape. Large and branched molecules more likely collide with the drift gas and hence have lower ion mobility than small molecules. The separation of the molecules depends on many factors including the length of the drift tube, temperature, drift gas, ambient pressure, and water content in the drift tube.

For VOC measurements in the life sciences it would be preferable if water saturated air could be used to deliver the analytes of

interest, like envisaged in breath research, were the patient blows directly into the equipment. To achieve maximal separation and at the same time sample preparation like water removal, the use of a multi-capillary column (MCC) in front of the IMS is powerful, increasing significantly the resolution of metabolites [23]. An MCC consists of approx. 1000 parallel, short capillaries that can withstand high gas flow rates of up to 250 mL/min (compared to 2 mL/min in a capillary GC column) resulting in short separation times. Notably, the MCC is insensitive to water and can be run with fully water saturated samples [25], thus allowing direct measurements of the fermentation off-gas without sample preparation. The water is not interacting with the column material and hence is flushed out before any analyte leaves the column. The total analysis time of one sample in an MCC-IMS is usually less than 500 s [26]. These characteristics together with the high sensitivity perfectly suit the MCC-IMS for online measurements of dilute volatile metabolites in the headspace of microbial fermentations.

The result is a heat map (Fig. 2) that shows the two-dimensionality of the MCC-IMS analysis. On the x-axis, the reduced inverse ion mobility from the IMS is given, while on the

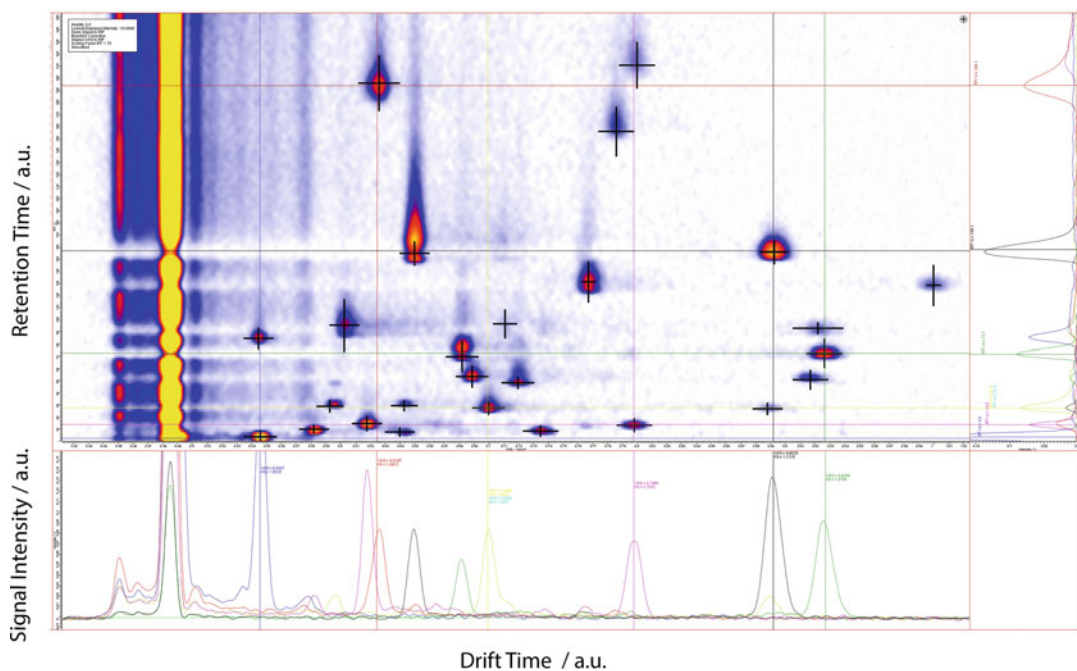


Fig. 2 MCC-IMS heat map. The *yellow* column is the reaction ion peak (RIP), indicating the amount of reaction ion molecules originating from the drift gas, here nitrogen. The *color code* indicates the voltage measured by the Faraday plate, which is proportional to the number of ions of a particular analyte, with *blue* very low and *yellow* very high. In addition, the chromatograms of the signal intensity (V) in dependence of the reduced inverse ion mobility and the retention time are given

y-axis the retention time from the MCC is given. The color code indicates the voltage measured by the Faraday plate, which is proportional to the number of ions of a particular analyte. The single parameters of reduced inverse ion mobility and retention time versus the voltage measured are given in separate plots on the bottom and the right of Fig. 2. The yellow column is the reaction ion peak (RIP), indicating the amount of reaction ion molecules originating from the drift gas, here nitrogen and residual water. The RIP signal is inverse proportional to the number of analyte ions, as for any molecule ionized one (or more) ion(s) of the RIP is (are) consumed. When a particular analyte is in too high concentrations or some high concentrated analytes co-elute from the MCC, the RIP reduces (*see* horizontal black line in Fig. 2) or if the concentrations are too high, disappears (not shown). The maximal analyte concentration detectable by an IMS is equal to the ions (here nitrogen) in the RIP. For further reading on particulars of IMS in general and drift tube IMS in particular, the reader is referred to the following IMS reviews [23, 27, 28].

1.3 Special Requirements for the Use of MCC-IMS for Microbial VOC Analysis

Two kinds of ^{63}Ni ion sources with activities of 95 MBq and 550 MBq are installed in commercial IMS systems, with the stronger one providing an enlarged dynamic range. In our research the two systems showed similar performance. The spectrometer with the weaker ionization source can be used in any laboratory, while operation of the system with the stronger ion source, e.g., in Germany, requires a handling permit of a state authority. Also, in Germany, the operator has to restrict access to and use of the instrument to trained persons and guarantee the secure removal of the potentially radioactively loaded off-gas of the instrument, e.g. via a fume hood. Alternative, radioactive ion sources, e.g., UV-ionization or MALDI, are available. In case of rather humid samples like breath or fermentation off-gas, spectrometers using UV ionization should not be used as these conditions lead to considerable reduction of the lifetime of the lamp.

To circumvent overloading of the highly sensitive IMS, the analytes might have to be diluted or a smaller sampling loop in front of the MCC has to be used. In analysis of VOCs emitted by yeast cultures, replacing the standard 100 mL loop with a 25 μL resulted in improved analyte separation. Also, the rather polar microbial VOCs might require alternative MCCs of different material as the standard MCC-IMS applications, e.g. breath analysis, deal with VOCs with different properties, might be suboptimal for the separation of microbial VOCs.

2 Materials

1. Microorganism to be investigated. We used *Saccharomyces cerevisiae* CEN.PK 113-7D, but the protocol can be used for any other microorganism.
2. Fermentation device: Fermenter or shake flask with metal cap.
3. Incubator shaker.
4. Growth medium appropriate for the respective microorganism (*see Note 1*).
5. Ultrapure water (*see Note 2*).
6. Off-gas tube made of polytetrafluoroethylene (PTFE), perfluoroalkoxy alkane (PFA) or Polyether ether ketone (PEEK) (*see Note 3*).
7. Ion mobility spectrometer: BreathDiscovery (B&S Analytik, Dortmund, Germany) (*see Note 4*).
8. Optional: Ventilated housing for the ion mobility spectrometer (*see Note 5*).
9. Multi-capillary column: S2-40/OV-1701/0.6 (length: 20 cm, Multichrom Ltd., Novosibirsk, Russia) (*see Notes 6 and 7*).
10. Measurement software: VOCan 3.6 (B&S Analytik, Dortmund, Germany).
11. Data evaluation software: VisualNow 3.5 (B&S Analytik, Dortmund, Germany).
12. PC or laptop with 4GB of RAM and Windows 7 or higher.
13. 100 mL glass bottle (*see Note 8*).
14. Carrier gas: Nitrogen 5.0 (purity: 99.999%) (*see Notes 9 and 10*).
15. Stainless steel pressure reducing unit and Swagelok connectors 1/8" for the gas connection from the gas bottle to the device (*see Note 11*).
16. Bottle multi-dispenser made of inert material such as PTFE, commercially available at laboratory supplier, e.g. Carl Roth, or steel (self-made, Fig. 3).

3 Methods

We describe here a basic protocol for the analysis of microbial volatile metabolites from set-up of the device to data acquisition and evaluation, and explain the sampling from the headspace of shake flask and bioreactor cultures and from the fermentation off-gas. While these set-ups only allow very low throughput, we argue that the MCC-IMS can easily be adapted to allow multiplexed

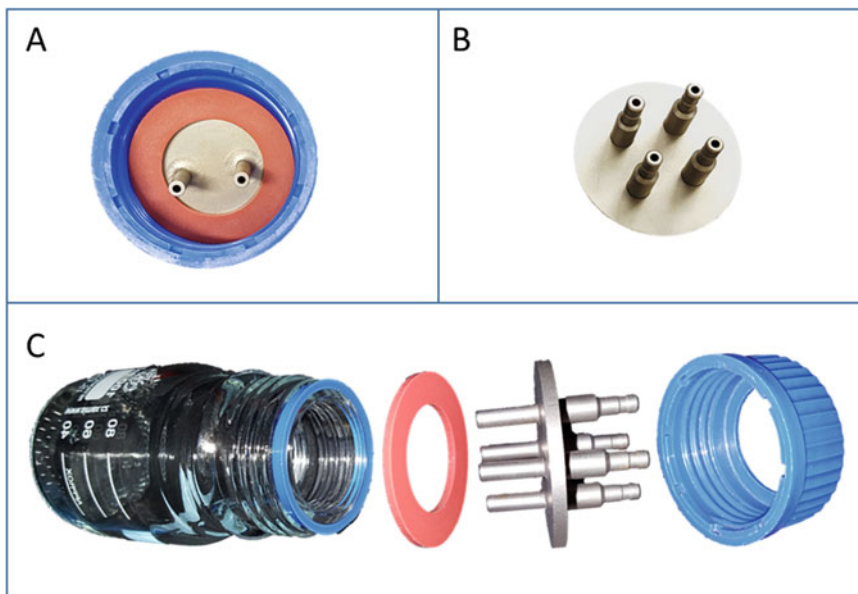


Fig. 3 Self-made multi-dispensers made of steel: Assembled two-way dispenser with rubber sealing placed in a screw cap (a), four-way dispenser (b). Assembly of four-way dispenser with rubber sealing (c)

sampling, e.g. from microtiter plates. A manual MCC-IMS data evaluation process is presented here, but in the IMS community software for automated peak detection and integration are underway.

1. Connect the MCC-IMS with the carrier and drift gas (here nitrogen).
2. In Germany: Make sure the outlet of the IMS is led outside via the chimney of the fume hood.
3. Always flush the device with carrier gas, even if not in use.
4. If the device was turned off, or was without carrier gas for a longer period of time, flush it until the RIP reaches a signal, in case of BreathDiscovery, in the range between 4 and 4.9 V to obtain a good sensitivity.

3.1 Set-Up of the Measurement Stand for Fermentation Off-Gas/Headspace Analysis

Sampling for MCC-IMS analysis is based on a sample loop. When not in use, the sampling tube is constantly flushed with nitrogen (by default 100 mL/min). During sampling, the nitrogen flow is stopped and a pump pulls air into the device (by default 10 s). In this time the sample is flushed through a sampling loop that has a constant volume of 10 mL. After loading for 10 s the content of the sampling loop is transferred onto the MCC-IMS, redirected towards the MCC-IMS. The sampling tube can be connected to virtually any container. We describe here how to sample from the off-gas or headspace of a fermenter, shake flask or glass bottle.

3.1.1 Sampling of Blank Air Sample to Check Device Function

Prior to the start of a measurement series, control measurements that check for proper function and the absence of device contamination are to be run. Use the set-up described below for these measurements. The same set-up is used to clean the device if any contaminations are detected (*see* Subheading 3.4 and **Note 12**).

1. Fill a clean 100 mL glass bottle with approx. 10 mL ultrapure water.
2. Connect the IMS sampling tube to the bottle using a multi-dispenser.
3. Adjust the sample tube in the glass bottle at a height that prevents that water is sucked into the MCC-IMS as the liquid might destroy the device.

3.1.2 Sampling from Shake Flask Head Space

Use the following set-up for headspace analysis of shake-flask experiments.

1. Insert the sample tube, connected to the MCC-IMS into the shake flask and close the flask with a metal cap. Use a sample tube made of PTFE or PFA tube as its stiffness ensures that the tube is squeezed but, not completely blocked when bent by closing the flask (*see* **Notes 13–16**).

3.1.3 Sampling from Bioreactor Off-Gas

1. Connect the uncooled off-gas outlet of the fermenter with a liquid trap (Fig. 4) to avoid carry-over of water droplets into the MCC-IMS (*see* **Notes 1** and **17**).
2. Connect the outlet of the liquid trap with the sample tube of the MCC-IMS.
3. In case, VOC concentrations in the sample lead to overloading of the IMS, dilute the off-gas with nitrogen fed into the liquid trap via the additional port of the dispenser (*see* **Note 18**).

3.1.4 Sampling from Bioreactor Head Space

1. Directly connect the sample tube to a sampling port of the bioreactor (*see* **Note 16**).

3.2 Set-Up of Measurement Programs

Measurement programs are used to perform automatic measurements and are set-up using the sequence editor built into VOCan. These programs consist of commands setting the gas flow rates, time of sampling, defining the sample ID, also comments for individual measurements can be included.

We describe the general procedure for the set-up of measurement programs and list the commands required for routine sample measurements and a device cleaning method in Table 1.

Prior to any sample measurement, two blank samples of humid (FN, in German, feuchte Null) and dry gas (TN, in German, trockene Null) are analyzed. In case the FN or TN measurements repeatedly show peaks not present in a clean device, the cleaning

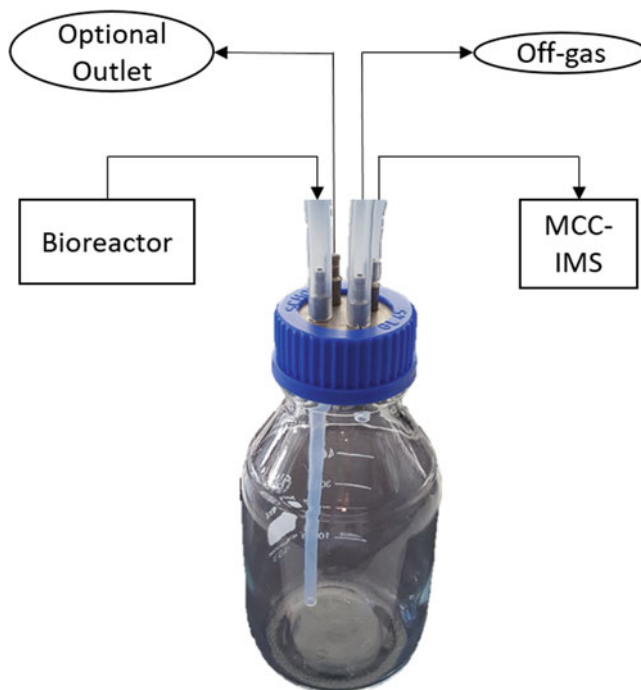


Fig. 4 Mixing flask with a four-way dispenser. The flask is connected to the bioreactor, with a long inlet tube, for the outlet to the MCC-IMS connection a short tube is used, for the off-gas no tube is required

Table 1
Measurement programs for standard MCC-IMS analyses

| # | Command | Description |
|---|--|--|
| Program for standard MCC-IMS measurements | | |
| 1 | flow sample 100 | Sample flow rate in [mL/min], default value is 100. |
| 2 | flow mcc 60 | Carrier gas flow rate through the MCC in [mL/min], default value is 60. |
| 3 | flow drift 100 | Drift gas flow rate in [mL/min], default value is 100. |
| 4 | wait with progress 30000 30 seconds pause | Pause of 30 s to allow for stabilization of flow rates. |
| 5 | sample id YYYY-MM-DD- EXPERIMENT NAME FN | Definition of the sample ID that is used as folder of the FN measurement. |
| 6 | comment | Optional input of comment, to be typed after the two slashes. |
| 7 | Prog: FN | Calls a predefined measurement program defined after the colon, here program FN. |
| 8 | wait with progress 10000 10 seconds pause | Pause of 10 s to stabilize the system after the measurement. |

(continued)

Table 1
(continued)

| # | Command | Description |
|----|--|---|
| 9 | sample id YYYY-MM-DD- EXPERIMENT NAME TN | Definition of file path of the TN measurement. |
| 10 | comment | Optional comment. |
| 11 | Prog: TN | Calls program TN. |
| 12 | wait with progress 10000 10 seconds pause | Pause of 10 s to stabilize the system after the measurement. |
| 13 | sample id YYYY-MM-DD- EXPERIMENT NAME M | Definition of file path and file name of the sample measurement. |
| 14 | comment | Optional comment. |
| 15 | Prog: Pump | Calls program “Pump”, which activates the sampling pump for 10 s to fill the sample loop. |
| 16 | flow sample 100 | Sample flow rate in [mL/min], set to the default value of 100. |
| 17 | flow mcc 150 | Carrier gas flow rate through the MCC in [mL/min], set to 150. |
| 18 | flow drift 100 | Drift gas flow rate in [mL/min], set to the default value of 100. |
| 19 | wait with progress 900000 15 minutes pause | Flushes the device for 15 min to clean it before the next measurement starts. |
| | Program for cleaning routine | |
| 1 | flow sample 100 | Sample flow rate in [mL/min], default value is 100. |
| 2 | flow mcc 60 | Carrier gas flow rate through the MCC in [mL/min], default value is 60. |
| 3 | flow drift 100 | Drift gas flow rate in [mL/min], default value is 100. |
| 4 | wait with progress 30000 Pause for 30 seconds. | Pause of 30 s to allow for stabilization of flow rates |
| 5 | sample id YYYY-MM-DD Clean FN | Definition of file path of the FN measurement |
| 6 | Prog: FN | Calls program FN. |
| 7 | wait with progress 10000 Pause for 10 seconds. | Pause of 10 s to stabilize the system after the measurement. |
| 8 | sample id YYYY-MM-DD Clean TN | Definition of file path of the TN measurement. |
| 9 | Prog: TN | Calls program TN. |
| 10 | wait with progress 10000 Pause for 10 seconds. | Pause of 10 s to stabilize the system after the measurement. |
| 11 | sample id YYYY-MM-DD Clean M | Definition of file path of the M measurement. |

(continued)

Table 1
(continued)

| # | Command | Description |
|----|--|---|
| 12 | Prog: Pump | Calls program “Pump”, which activates the sampling pump for 10 s to fill the sample loop. |
| 13 | wait with progress 10000 Pause for 10 seconds. | Pause of 10 s to stabilize the system after the measurement. |
| 14 | flow sample 300 | Sample flow rate in [mL/min], set to 300. |
| 15 | flow mcc 250 | Carrier gas flow rate through the MCC in [mL/min], set to 250. |
| 16 | flow drift 300 | Drift gas flow rate in [mL/min], set to 300. |
| 17 | wait with progress 10800000 Pause for 3 hours. | Flushes the device with high flow rates as defined above for 3 h (<i>see Note 19</i>). |

routine is to be run, which basically consists of repetitive measurements of ultrapure water, which gradually removes the contamination. This might take between few hours and several days, depending on the kind and amount of the contamination. Another indication for a clean device is the RIP intensity, which should be in the range from 4 to 4.9 V, but is reduced in a dirty device. Yet another indication for a clean device, is to record an entry curve.

1. Open VOCan.
2. Right-click in the sequence field and choose editor.
3. Type or paste the commands for sample measurements or device cleaning given in Table 1 into the sequence field.
4. Save the sequence by right-click into the sequence field and selection of save sequence. Choose save as .seq file.
5. Save the program for standard measurements under the name “Standard measurements” and the program for the cleaning procedure under “Cleaning routine”.

3.3 Check the MCC-IMS for Contamination

Prior to any sample measurement run a control measurement with water saturated air of ultrapure water to check for any contaminations or malfunctioning of the device.

1. Connect the MCC-IMS sampling tube with the water bottle as described in Subheading 3.1.1.
2. Open VOCan.
3. Press [Ctrl + m] to switch to a user interface with more options (*see Fig. 5*).
4. Click on the Settings tab in the upper right corner (*see Fig. 5c*) and check the box use sample ID for save path (*see Note 20*).

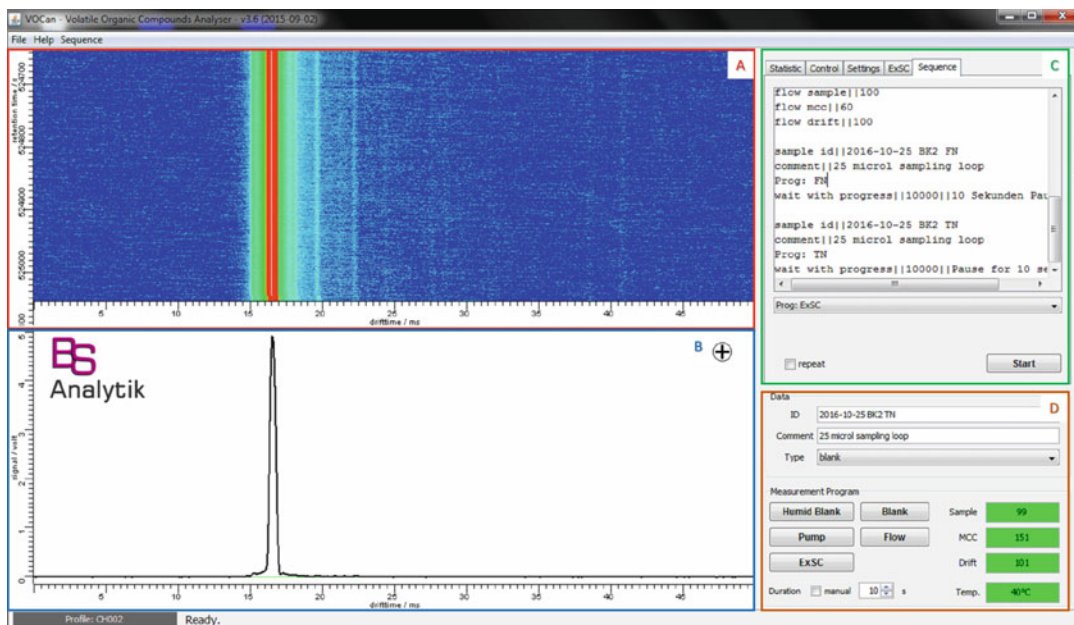


Fig. 5 User interface of VOCan. A heat plot of the latest recorded spectra is displayed in (a). The mean of the latest recorded spectra is displayed in (b), along with a plus or a minus for positive-, respectively, negative-ion mode. The settings available in VOCan can be entered in (c). (d) shows a control panel for starting single measurements and the current status of flow values

5. Adjust the length of the measurement under Measurement duration to a value larger than the longest retention time of any detected peak. The default value is 750 s.
6. Check if the gas flow rates (sample, MCC and drift gas) and the temperature are at their preset values. The actual values are given in the lower right corner and are highlighted in green in the accepted range (Fig. 5d).
7. Load the measurement program Cleaning routine by right-click into the sequence window and selection of the respective file.
8. Start the measurement program by clicking the Start button (Fig. 5d).
9. When the measurement is completed, load the measurement set into VisuaNow following the **steps 2–5** described in Subheading 3.7.
10. Compare the 3D topographic plot of the control measurement (cf. Subheading 3.3) and the FN measurement in VisualNow. Follow the steps described in Subheading 3.5. If the contamination is detected in both the M measurement and in the FN measurement, the device is dirty. If the contamination only occurs in the M measurement, most likely the sample tube is contaminated and should be exchanged. The plots should

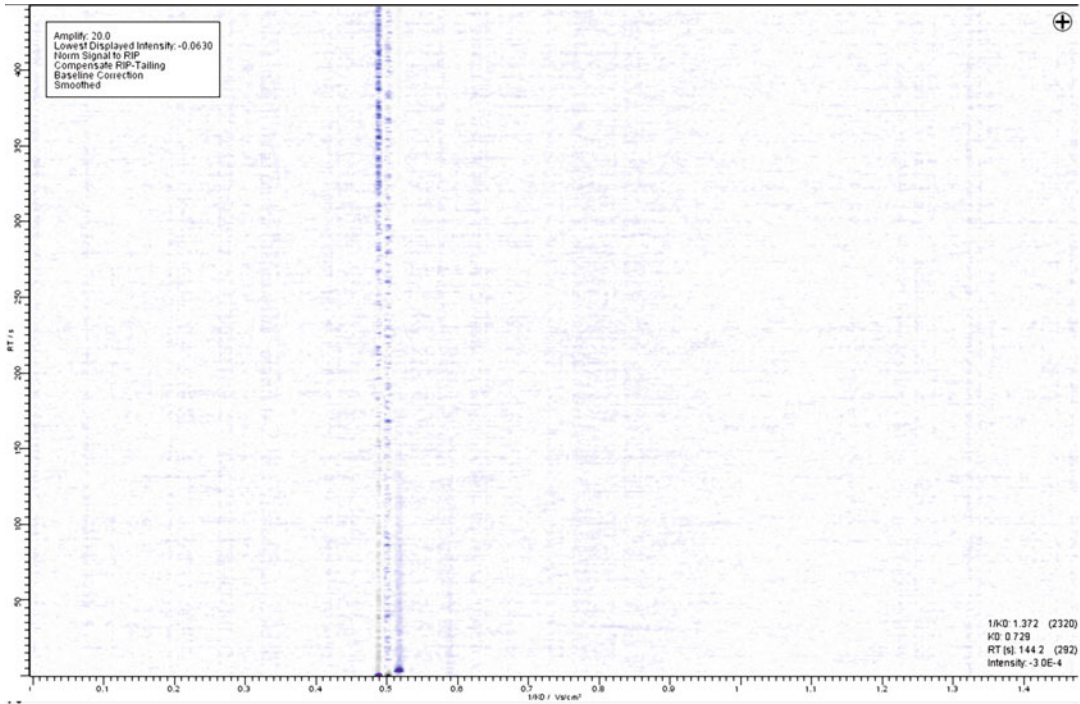


Fig. 6 Topographic plot of a dry blank sample in a properly functioning and clean MCC-IMS

display background noise and the RIP only (cf. Fig. 6). If any peaks of significant intensity are detected, perform the steps described in Subheading 3.4. If the device is clean directly start the sample measurements (cf. Subheading 3.5).

3.4 Cleaning of a Contaminated Device

1. Keep the MCC-IMS attached to the water bottle (cf Subheading 3.3).
2. Check the box repeat below the sequence window in VOlCan (Fig. 5c).
3. Load the measurement program Cleaning routine by right-click into the sequence window and selection of the respective file.
4. Start the program by clicking the Start button (Fig. 5d).
5. Run the cleaning procedure overnight or even for a couple of days until the contamination has disappeared.

3.5 MCC-IMS Measurements

1. Connect the MCC-IMS sampling tube with the shake flask or fermenter (cf. Subheadings 3.1.2, 3.1.3 or 3.1.4.).
2. Perform **step 2–6** of Subheading 3.3).
3. Load the program “Standard measurements” by right-clicking into the sequence window and selecting the respective file.
4. Start the program by clicking the Start button (Fig. 5d).

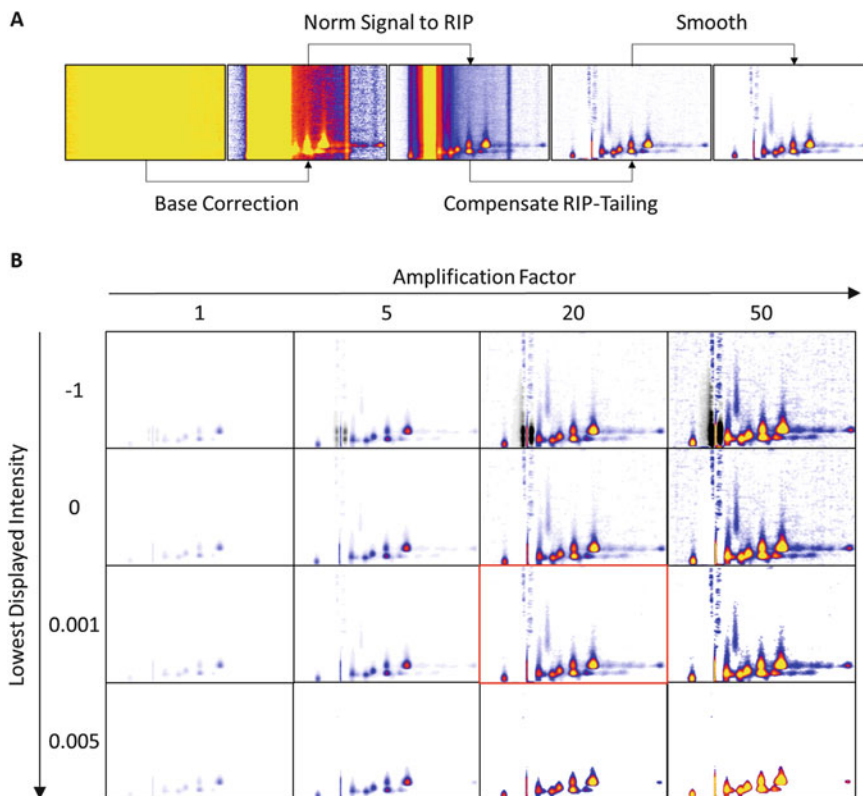


Fig. 7 Influence of different correction and smoothing parameters (a) and the amplification factor and lowest display intensity (b) on the heat plot in VisualNow

3.6 Configuration of the Data Evaluation Software Visual Now

1. Open VisualNow.
2. Go to Parameters in the menu bar and make sure Base Correction, Norm Signal to RIP, Compensate RIP-Tailing, Smooth, and Median Smooth are ticked. With these settings, peaks become visually analyzable (see Fig. 7a).
3. It is important that the scaling factor is set to 1 and alignments are turned off as the Peak Analyst (Subheading 3.7, step 19) will ignore the scaling factor and peaks will have another retention time in the Peak Analyst compared to the scaled data in the main window of VisualNow.

3.7 Data Evaluation and Interpretation

With the following protocol, intensities of peaks, defined in the layer are extracted from the measurement data set and exported in an Excel file. Also, series of excerpts of the topographic plot displaying the peak are generated and the data exported into an Excel or .png file.

1. Open VisualNow (see Note 21).
2. Load a .csv file by selecting it from the Tree tab (see Notes 21 and 22) (Fig. 8).

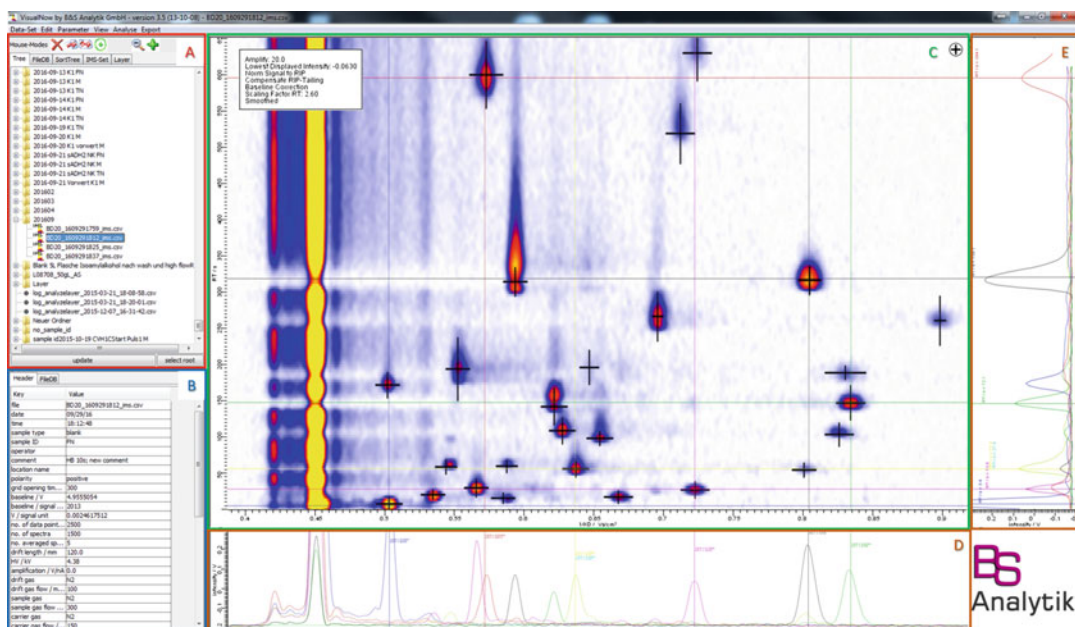







Fig. 8 Graphical user interface of VisualNow. The tool window (a) provides functions for peak annotation and addition of files to a set and tabs for file selection and layer adaptation. The box in the lower left corner (b) shows additional information on the selected file. A topographic plot of the selected data is shown in the center (c) with the IMS spectrum for the selected retention time (depicted by the *cross line*) below (d) and the MCC chromatogram for the selected 1/KO value (depicted by the *cross line*) on the right (e)

3. Inspect all humid blank (FN) and dry blank (TN) files for possible contaminations of the system. To zoom into the topographic plot, left-click and hold mouse key to drag a window around the region that shall be enlarged.
4. If there is no contamination, load all measurement (M) files of the series into a set by selecting them and clicking the “add selected data” to IMS-Set button (+) once (see Fig. 8a and Note 23).
5. To make measurement of distinct experiments comparable, it is advised to scale retention time and reduced inverse ion mobility to data obtained from a mixture of reference analytes with well-known concentration.
6. Click on IMS-Set in the Tab menu (Fig. 8a).
7. Click the “for adding peaks to layer” button (+) (Fig. 8a).
8. Load the first file of the set by switching to the IMS-Set tab in the upper left corner and selecting the file.
9. Check the topographic plot for potential peaks and pick them by left-click (see Note 24). The peak will be marked by a small black cross and will be automatically assigned a peak name (syntax P#, with # being a consecutive number). The cross

size can be changed with the Scale Annotations button () (left-click and drag) and the position can be adjusted by using the Move Annotations button () (Fig. 8a) (*see* **Notes 25 and 26**). At this step a coarse fit is sufficient as the peak annotation will be optimized in Subheading 3.7, **step 14–23**. To delete a peak, select the Delete Annotations button () and click on the annotation you want to delete (*see* **Note 27**).

10. The IMS spectrum and MCC chromatograms below and on the left side of the topographic plot (Fig. 8d, e) can also be used for peak verification. To do so, deselect the current annotation tool (Subheading 3.7, **step 7 or 9**) and simply left-click on the peak to be investigated, a black crosshair will appear and the spectrum and chromatogram of the respective retention time and inverse ion mobility are displayed. Now check the chromatograms for peaks. A peak ideally looks symmetrical, and is clearly distinct of the baseline. If the peak is not symmetrical and tailing on the y-axis, this indicates an overloaded MCC. If the peak has a shoulder, or a second maximum, this indicates a second peaks that overlaps, with the investigated peak. An investigation of the other axis (y-axis for retention time or x-axis for $1/K_0$) can help to investigate overlapping peaks in greater detail.
11. Repeat **steps 6–9** for all files in the measurement set (*see* **Note 28**).
12. Click on the Layer tab (Fig. 8a) and select user peaks (now highlighted in yellow).
13. Save the layer that contains all peak annotations, using the Save Layer as .xls button (). This saves the peak annotations that were created in **steps 6–11**. If the program crashes, the layer can be restored by using the Load Layer from .xls or .pl button ()
14. Select Analyze (menu bar) and click on Peak Analyst.
15. A new window opens, select the first peak and click on show regions (bottom).
16. Select update (Fig. 9d). This may take a while for the first peak but is considerably quicker for the following peaks. Select update spectra (Fig. 9d) (*see* **Note 29**).
17. An excerpt of the heatmap around the selected peak of all files is displayed (Fig. 9b). Below the heat maps, the IMS spectra of the selected peak in all files, are plotted (Fig. 9c).
18. Adjust the amplification factor and the lowest displayed intensity to optimize the topographic plot. Carefully adjust the amplification factor to not attenuate weak signals below visibility or to increase the intensity too much, which will result in overlapping peaks. The same holds for the lowest displayed intensity as a too low value will result in misannotation of background noise as peaks, while a too high value might lead to loss of peaks (Fig. 7b).

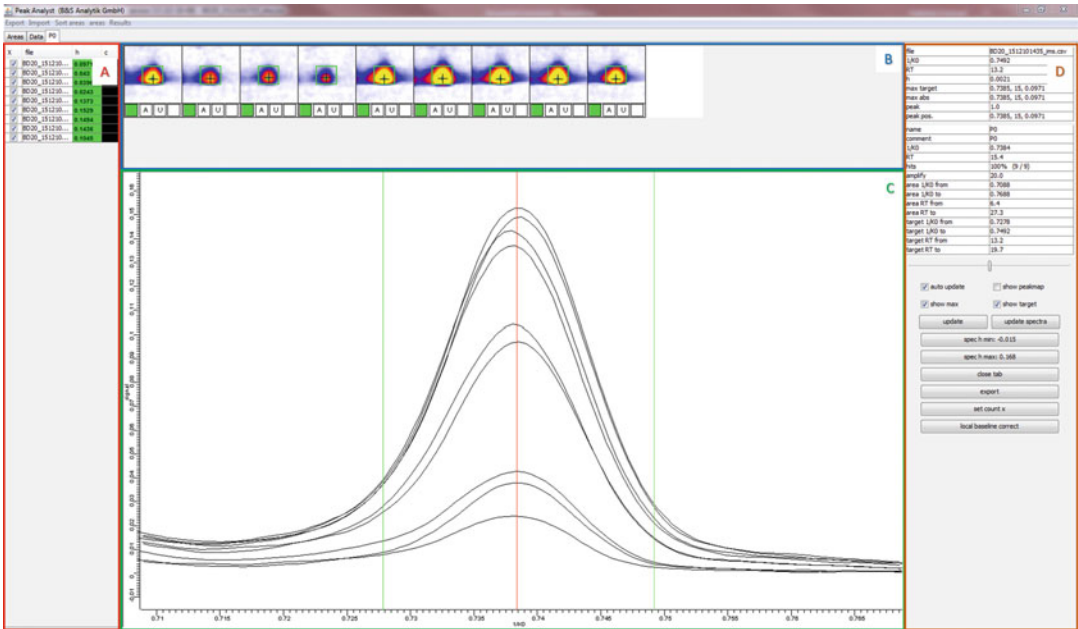


Fig. 9 Peak Analyst window. (a) Files present in dataset. (b) Excerpt of the topographic plots of the single measurements in the dataset showing the annotated peak P0. (c) IMS spectra of peak P0, determined by the green rectangle in B, in all selected files. (d) Additional information on the peak as well as control elements

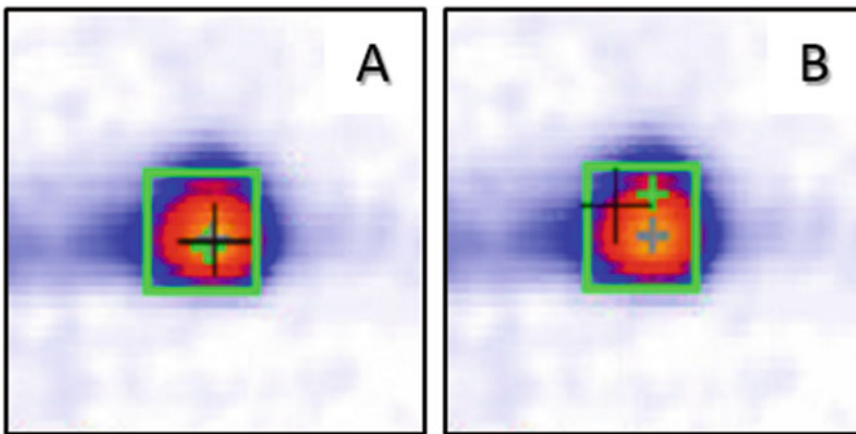


Fig. 10 Zoomed-in plot of a peak in the Peak Analyst. (a) Ideal peak annotation. (b) Erroneous peak annotation

19. Visually inspect the single plots in the Peak Analyst window (Fig. 9b) and correct the peak annotation if necessary. Ideally, the different cross lines lie on top of each other and the green rectangle comprises the peak region (Fig. 10a). The green rectangle enclosed the area used to determine the maximum intensity of the peak. To adjust the peak area, draw a green rectangle by right mouse click and drag the mouse in one direction, this will change the area of the detected peak. The

rectangle should cover the whole area of the peak, but should not be bigger than necessary. It is recommended to use a fixed amplification factor for all of the peaks and choose an area that spans the red or yellow region of the peaks (*see Note 30*). The black cross indicates the local maximum for each file. It should be located on the maximum of the peak and not on the maximum of neighboring peaks. The cross can be moved manually by left-clicking on the maximum of the correct peak. The green cross indicates the average local maximum intensity within the green rectangle for all files in the set, while the grey cross lines indicate the average maximum of the complete, depicted area (not green rectangle) over all files. The maximal intensity inside the green rectangle is used as the intensity value of the peak in Subheading 3.7, **step 28 and 29**. Fig. 10b depicts a situation that requires manual interpolation.

20. Click again on update spectra.
21. Check the plot of the IMS spectra (Fig. 9c). The green lines should enclose the peak and the red line should be where most of the spectra have their maximum. If this is not the case the green rectangle drawn in **step 19** has to be readjusted until the green lines enclose the peak of the spectra, the ideal case is depicted in Fig. 7c. To move the red line in Fig. 7c, manual interposition is necessary. This can be done by moving the black cross as described in **step 19**.
22. When the peak annotation is optimized, go back to the Areas tab (**step 15**) and redo **steps 16–21** for all other peaks defined in the layer (*see Notes 31 and 32*).
23. Click on Export in the menu bar and choose Export complete and specify the file path. It is recommended to create a new folder since several files will be generated. Run the statistics- and class-separation analysis with the option “choose horizontal and vertical tiles count”.
24. Minimize the Peak Analyst window and go back to the Layer tab in VisualNow. To load the targets_layer from the folder that you exported to in **step 23**, click in the Layer tab on the Load Layer from .xls or .pl button (📄). After that click on the X left to user peaks, making the layer disappear.
25. Look at the layer (displayed in Fig. 8c) and compare if there are rectangles at the same position or overlapping. If necessary, use the Peak Analyst to erase double annotations (over one peak) or reduce peak overlapping. If you need to adapt the target layer, delete it in the Layer tab by clicking on the Remove/Clear Layer button (✖) and proceed by deleting the content of the folder you created in **step 23**. Proceed with **step 23** until no further changes are necessary.
26. Make sure the user peaks target layer is selected (highlighted in yellow), and if it is not, do so by left-clicking on it. To obtain a graphical representation of all peaks over the time of the

measurement choose Analyze from the menu bar and click on Compare Peak-Images from Layer.

27. Save the image by clicking on Export Image at the bottom of the page and close the window. The image shows a heat plot of each peak on the X-axis, while the files from your layer are plotted on the Y-Axis.
28. To obtain values for the peaks choose again Analyze from the menu bar and click on Observe data directory.
29. Check the boxes include all files from set and analyze from layer. Choose user peaks targets from the dropdown menu and uncheck the box add to Set. Then, press start.
30. Two windows will open; the left window shows peak positions and the image of the peak from the last file. The right window shows the intensity of the different peaks over time. It can be used to easily display the different peaks by selecting them using the left y-axis and right y-axis tabs in the menu bar. To zoom in simply draw a rectangle by holding the left mouse button to zoom out, right-click and select Autojustage, then Beide Achsen (Engl., both axes). Use the Export tab in the menu bar and select As .xls, which will save the peak intensities in an Excel file. In this Excel sheet, the heat plots of the different peaks sorted column-wise with heat plots of the single measurements in the single rows (*see Note 33*).

3.8 Reuse of Peak Annotation and Evaluation Files

1. For datasets recorded under similar conditions reuse the same target layer without processing the data evaluation procedure (cf Subheading 3.7, **step 24**) for every single dataset.
2. Load all files of the new dataset into the IMS-Set, as described in Subheading 3.7, **step 4** and simply load the target_layer in the layer tab of VisualNow (cf. Subheading 3.7, **step 23**)
3. Browse the dataset and check for peaks without a target area. If there are no peaks without target area proceed with Subheading 3.7, **step 26**. If there are peaks without a target area, proceed with **step 4**.
4. Select all new peaks as described in Subheading 3.7, **steps 6–10** and select user peaks in the Layer tab in VisualNow.
5. Choose Analyze from the menu-bar and open the Peak Analyst.
6. Select Open on the lower side of the window. Open the .peakAn file from a previous data evaluation.
7. Now all peaks from the previous data evaluation are loaded in the peak analyst. To load the newly selected peaks, select Import in the menu bar on top of the window and choose Import areas from layer.
8. Proceed with the use of the Peak Analyst as described from Subheading 3.7, **step 15** onwards.

3.9 Preparation of an Entry Curve

This is an additional, optional, method to check the functionality of the instrument.

1. Measure a sequence of ultrapure water as described in Subheading 3.4. It is recommended to run about 50 measurements.
2. Load the files into an IMS-set as described in Subheading 3.7, step 4.
3. Turn off Compensate RIP-Tailing by choosing Parameters in the menu bar and deactivating Compensate RIP-Tailing.
4. Create about 5 windows on the RIP, by using a right-click and dragging a window (a window with a letter will appear on the screen), that is completely located in the RIP area.
5. Choose Analyze from the menu bar and then Observe data directory. Check include all files from set and analyze from layer. Uncheck add to Set. Finally, hit Start.
6. Compare the intensities over time. In a clean device, the displayed intensities will remain constant over time.

4 Notes

1. We recommend working with minimal salt medium as complex medium might result in high background in the MCC-IMS measurements due to volatile medium components and possible interference with target compounds. For example, we detected more than 100 peaks in measurements of non-inoculated industrial molasses medium.
2. We used water filtered with PURELAB flex (ELGA, Celle, Germany) with a resistivity of maximal 18.2 M Ω -cm as deionized water sometimes led to ghost peaks.
3. This material is gas-tight and was tested not to contain compounds that cause peaks in the IMS. In addition, no or minimal analyte adsorption to the inner walls of the tubes was reported.
4. We used the model BD20, a new product line of the Breath-Discovery is now available.
5. Because of the high sensitivity of the IMS, we experienced ghost peaks potentially caused by VOCs present in the air of the fermentation lab (e.g., solvents or released from autoclaved microbial cultures). To avoid such interferences, we encapsulated the device in a Plexiglas housing with inlets and outlets for the required gases and the samples. The housing was continuously flushed with compressed air and efficiently shielded the IMS from the exterior air; the temperature in the housing stayed constant, at around ambient.

6. MCCs are using the same separation principle, as gas chromatographic columns, but instead of one long capillary consist of approx. 1000 parallel columns. This allows higher gas flows through the column and thus a faster separation compared to standards GC columns.
7. The standard MCC used in the BreathDiscovery is the MCC S2-40/OV-5/0.2 (polydimethylsiloxane (95%); diphenyl (5%); film thickness 0.2 μm). We used the OV-1701 (polydimethylsiloxane (86%); cyanopropylphenyl (5%); film thickness 0.6 μm) because it showed better separation of polar yeast volatile metabolites, resulting in more evenly distribution of the peaks on the y-axis in the topographic plots. By better separation, the amount of molecules entering the IMS at one given time point is reduced. This is accompanied by less competition for ionization and reduced risk of overloading of the device. Note, that the column can only be exchanged by a technician from B&S Analytik GmbH.
8. To ensure that the glass bottle is clean and odor-free, connect the bottle with the sample tube of the MCC-IMS and flush the bottle with carrier gas from prior to the measurements.
9. Alternatively, synthetic air can be used. For example, Air Liquide produces synthetic air in IMS quality. High purity of the gas flowing into the IMS is necessary to maximize analyte signals. We used nitrogen 5.0 with an additional hydrocarbon gas filter.
10. Pressure reducers made of brass should be avoided, they tend to lead to permanent reduction of the RIP intensity due to impurities potentially released from the brass material. It is also possible that the RIP signal is not permanently reduced, but only in irregular time intervals. In the best case the complete gas supply is based on PTFE, PPEK or PFA tubing and stainless steel that was flushed prior to the usage with clean nitrogen or synthetic air.
11. All gas lines should be flushed for a minimum of 10 min with clean gas before the BreathDiscovery is connected.
12. Some contamination might not appear in the heatmaps of dry samples. In that case the contaminations will not be detected, but will also not be washed out of the system, that is why a humid sample is necessary.
13. The main yeast culture can either be shaken, or be held stationary for the time of the experiment.
14. In between measurements, the sampling tube is back-flushed with the carrier gas. If the cultivation shall be done under fully aerobic conditions, use synthetic air as carrier gas, or actively aerate the shake flask with a second gas inlet, or remove the sampling tube in between the measurements.

15. If the shake flask shall be placed in an incubator shaker during the measurement, make sure the sample tube connecting the flask with the BreathDiscovery is long enough to allow rotation of the flask. Fix the tube at the housing or door of the shaker. Strictly avoid that fermentation broth enters the sample tubing since liquid can destroy the IMS.
16. In this set-up the drift-gas will be pumped into the bioreactor in between the measurements. While this avoids accumulation of condensed water and VOCs in the sampling line, it comes with the risk of culture contamination.
17. Ideally, the complete off-gas and sampling line is heated to fermentation temperature, to avoid water or VOC condensation. Keep in mind, that the heated gas should not be far above the MCC temperature (default 40 °C) to avoid disturbances of the column temperature at column inlet.
18. Given the limited dynamic range of the IMS, the bioreactor off-gas might have to be diluted by a factor of 1:100 or more.
19. The pause time can be adjusted by the user. If analytes are eluting from the column after the measurement in VOCan at retention times larger than the measurement time the measurement duration needs to be increased, since there are still compounds eluting from the MCC, that could interfere with new measurements. However, the measurement duration can be decreased, if the last compound elutes significantly before the end of a measurement.
20. With this box checked, the sample ID will create a new folder and automatically sort the different types of measurements in different folders. The box can also be left unchecked, however, in this case, we recommend a logbook entry for every measurement.
21. If VisualNow is already open, click the update button below the tree window.
22. If there are no files in the Tree tab, use the select root button at the lower right corner of Fig. 8a.
23. For batch loading of several data files select the first file by mouse click, hold the [Shift] key on your keyboard and select the last file to be analyzed. Selection and deselection of single files is possible by mouse click on the file name while pressing the [Ctrl] key.
24. Due to diffusion effects in the MCC, peaks broaden with increasing retention times. This is already considered by the peak tool (*see* Subheading 3.7, step 9).
25. In case peaks with low intensity cannot unambiguously be distinguished from background noise, increase the signal amplification parameter. This parameter can be changed in the Parameter menu. To amplify peak signals set the value to

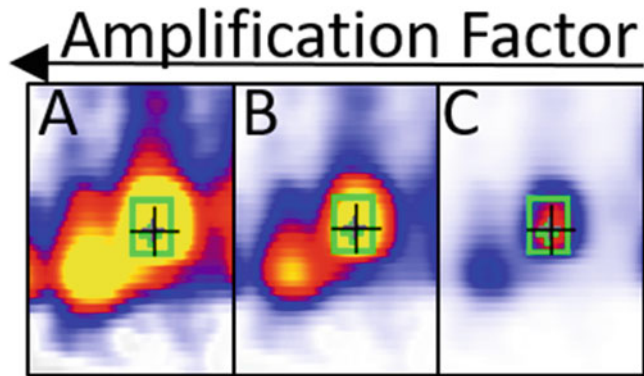


Fig. 11 Example for overlapping peaks. The green rectangle was chosen relatively small to prevent peak overlapping. **a**, **b**, and **c** determine the amplification factor of the same two peaks. **(a)** 5, **(b)** 20, **(c)** 50

50. Likewise, reduce amplification to distinguish overlapping peaks. Reset the parameter to the default value of 20 after peak evaluation.
26. There is no need to readjust the peak annotation in every single measurement file at this step as this fine-tuning is more easily done with the Peak Analyst tool (*see* Subheading 3.7, **step 14–23**).
27. All peak annotations will be saved as one layer, with which the whole measurement set will be evaluated. Therefore, peaks appearing in later datasets should be added but peaks only present in a subset should not be deleted.
28. Select all peaks including the blue ones, although of low intensity they may hold valuable information.
29. The number of files per line, default value is 5, can be altered by clicking on set count x on the right side. A dialog window will open, in which the number can be set.
30. Figure 11 shows an example of poorly separated peaks: The green rectangle marks the contour of the right peak, which is partially overlapped by a second peak. Reducing the amplification of the peak intensity reveals that the rectangle encompasses the overlapping region of the two peaks, which should be avoided to evaluate the maxima of the two overlapping peaks, without the interference of the other peak.
31. If there are double peaks, or peaks that do not show a (significant) amplitude in the spectrum, they can be deleted by selecting them in the Areas tab and clicking on remove area.
32. To rename peaks, go to the Areas tab, select the peak, change the entries of name and comment (Fig. 9d). End the editing process by pressing [Enter] on your keyboard. Finally click on the commit button below the name and comment section and choose the next peak to rename.

33. To extract the time from the file names a macro can be used. The filename (e.g., BD20_1609211735_ims.csv) is BD20_YYMMDDhhmm_XXX, that is, device name (BD20), date and time the file was created (YYMMDDhhmm), and additional information (XXX) required for the computer to read the file.

References

1. Petzold CJ, Chan LJ, Nhan M, Adams PD (2015) Analytics for metabolic engineering. *Front Bioeng Biotechnol* 3:135. doi:10.3389/fbioe.2015.00135
2. Billeci K, Suh C, Di Ioia T, Singh L, Abraham R, Baldwin A, Montecarlo S (2016) Implementation of an automated high-throughput plasmid DNA production pipeline. *J Lab Autom* 21(6):765–778. doi:10.1177/2211068216630547
3. Dietrich JA, McKee AE, Keasling JD (2010) High-throughput metabolic engineering: advances in small-molecule screening and selection. *Annu Rev Biochem* 79:563–590. doi:10.1146/annurev-biochem-062608-095938
4. Ukibe K, Katsuragi T, Tani Y, Takagi H (2008) Efficient screening for astaxanthin-overproducing mutants of the yeast *Xanthophyllomyces dendrorhous* by flow cytometry. *FEMS Microbiol Lett* 286(2):241–248. doi:10.1111/j.1574-6968.2008.01278.x
5. Kim SW, Keasling JD (2001) Metabolic engineering of the nonmevalonate isopentenyl diphosphate synthesis pathway in *Escherichia coli* enhances lycopene production. *Biotechnol Bioeng* 72(4):408–415
6. Olson ML, Johnson J, Carswell WF, Reyes LH, Senger RS, Kao KC (2016) Characterization of an evolved carotenoids hyper-producer of *Saccharomyces cerevisiae* through bioreactor parameter optimization and Raman spectroscopy. *J Ind Microbiol Biotechnol* 43(10):1355–1363. doi:10.1007/s10295-016-1808-9
7. Kreyenschulte D, Paciok E, Regestein L, Blümich B, Büchs J (2015) Online monitoring of fermentation processes via non-invasive low-field NMR. *Biotechnol Bioeng* 112(9):1810–1821. doi:10.1002/bit.25599
8. Luoma P, Golabgir A, Brandstetter M, Kasberger J, Herwig C (2016) Workflow for multi-analyte bioprocess monitoring demonstrated on inline NIR spectroscopy of *P. chrysogenum* fermentation. *Anal Bioanal Chem* 409(3):797–805. doi:10.1007/s00216-016-9918-9
9. Levin-Karp A, Barenholz U, Bareia T, Dayagi M, Zelcbuch L, Antonovsky N, Noor E, Milo R (2013) Quantifying translational coupling in *E. coli* synthetic operons using RBS modulation and fluorescent reporters. *ACS Synth Biol* 2(6):327–336. doi:10.1021/sb400002n
10. Mutalik VK, Nonaka G, Ades SE, Rhodius VA, Gross CA (2009) Promoter strength properties of the complete sigma E regulon of *Escherichia coli* and *Salmonella enterica*. *J Bacteriol* 191(23):7279–7287
11. Rhodius VA, Mutalik VK, Gross CA (2012) Predicting the strength of UP-elements and full-length *E. coli* sigmaE promoters. *Nucleic Acids Res* 40(7):2907–2924. doi:10.1093/nar/gkr1190
12. Zobel S, Benedetti I, Eisenbach L, de Lorenzo V, Wierckx N, Blank LM (2015) Tn7-based device for calibrated heterologous gene expression in *Pseudomonas putida*. *ACS Synth Biol* 4(12):1341–1351. doi:10.1021/acssynbio.5b00058
13. Binder S, Schendzielorz G, Stabler N, Krumbach K, Hoffmann K, Bott M, Eggeling L (2012) A high-throughput approach to identify genomic variants of bacterial metabolite producers at the single-cell level. *Genome Biol* 13(5):R40. doi:10.1186/gb-2012-13-5-r40
14. Dietrich JA, Shis DL, Alikhani A, Keasling JD (2013) Transcription factor-based screens and synthetic selections for microbial small-molecule biosynthesis. *ACS Synth Biol* 2(1):47–58. doi:10.1021/sb300091d
15. Siedler S, Stahlhut SG, Malla S, Maury J, Neves AR (2014) Novel biosensors based on flavonoid-responsive transcriptional regulators introduced into *Escherichia coli*. *Metab Eng* 21:2–8
16. Mustafi N, Grunberger A, Kohlheyer D, Bott M, Frunzke J (2012) The development and application of a single-cell biosensor for the detection of L-methionine and branched-chain amino acids. *Metab Eng* 14(4):449–457
17. Paige JS, Nguyen-Duc T, Song WJ, Jaffrey SR (2012) Fluorescence imaging of cellular

- metabolites with RNA. *Science* 335 (6073):1194–1194
18. Ebert BE, Halbfeld C, Blank LM (2016) Exploration and exploitation of the yeast volatilome. *Curr Metabolomics* 4:1–17. doi:[10.2174/2213235X04666160818151119](https://doi.org/10.2174/2213235X04666160818151119)
 19. Haggarty J, Burgess KE (2016) Recent advances in liquid and gas chromatography methodology for extending coverage of the metabolome. *Curr Opin Biotechnol* 43:77–85. doi:[10.1016/j.copbio.2016.09.006](https://doi.org/10.1016/j.copbio.2016.09.006)
 20. Wu C, Siems WF, Hill HH (2000) Secondary electrospray ionization ion mobility spectrometry/mass spectrometry of illicit drugs. *Anal Chem* 72(2):396–403. doi:[10.1021/ac9907235](https://doi.org/10.1021/ac9907235)
 21. Barrios-Collado C, García-Gómez D, Zenobi R, Vidal-de-Miguel G, Ibáñez AJ, Martínez-Lozano Sinues P (2016) Capturing *in vivo* plant metabolism by real-time analysis of low to high molecular weight volatiles. *Anal Chem* 88(4):2406–2412. doi:[10.1021/acs.analchem.5b04452](https://doi.org/10.1021/acs.analchem.5b04452)
 22. Fink T, Baumbach JI, Kreuer S (2014) Ion mobility spectrometry in breath research. *J Breath Res* 8(2):027104. doi:[10.1088/1752-7155/8/2/027104](https://doi.org/10.1088/1752-7155/8/2/027104)
 23. Baumbach JI, Eiceman GA (1999) Ion mobility spectrometry: arriving on site and moving beyond a low profile. *Appl Spectrosc* 53(9):338a–355a. doi:[10.1366/0003702991947847](https://doi.org/10.1366/0003702991947847)
 24. Halbfeld C, Ebert BE, Blank LM (2014) Multi-capillary column-ion mobility spectrometry of volatile metabolites emitted by *Saccharomyces cerevisiae*. *Meta* 4(3):751–774. doi:[10.3390/metabo4030751](https://doi.org/10.3390/metabo4030751)
 25. Xie Z, Sielemann S, Schmidt H, Baumbach JI (2000) A novel method for the detection of MTBE: ion mobility spectrometry coupled to multi capillary column. *Int J Ion Mobil Spectrom* 4:77–83
 26. Baumbach JI (2006) Process analysis using ion mobility spectrometry. *Anal Bioanal Chem* 384(5):1059–1070. doi:[10.1007/s00216-005-3397-8](https://doi.org/10.1007/s00216-005-3397-8)
 27. Cumeras R, Figueras E, Davis C, Baumbach JI, Gracia I (2015) Review on ion mobility spectrometry. Part 1: current instrumentation. *Analyst* 140(5):1376–1390. doi:[10.1039/c4an01100g](https://doi.org/10.1039/c4an01100g)
 28. Cumeras R, Figueras E, Davis C, Baumbach JI, Gracia I (2015) Review on ion mobility spectrometry. Part 2: hyphenated methods and effects of experimental parameters. *Analyst* 140(5):1391–1410. doi:[10.1039/c4an01101e](https://doi.org/10.1039/c4an01101e)

Selection of Highly Expressed Gene Variants in *Escherichia coli* Using Translationally Coupled Antibiotic Selection Markers

Maja Rennig, Daniel O. Daley, and Morten H.H. Nørholm

Abstract

Strategies to select highly expressed variants of a protein coding sequence are usually based on trial-and-error approaches, which are time-consuming and expensive. We address this problem using translationally coupled antibiotic resistance markers. The system requires that the target gene can be fused at the 3'-end with a translational coupling element and an antibiotic resistance gene. Highly expressed target genes can then be selected using a fast and simple whole cell survival assay in the presence of high antibiotic concentrations. Herein we show that the system can be used to select highly expressing clones from libraries sampling translation initiation sites.

Key words Gene expression, Protein production optimization, Selection, Library screening, Antibiotic resistance, Translational coupling

1 Introduction

Bacterial production of recombinant proteins is highly important in the construction of cell factories, and for basic studies on the function, interactions, and structure of proteins. Unfortunately yields are often low, particularly in heterologous hosts, and optimization of the coding sequence is necessary. Coding sequence optimizations frequently utilize randomized libraries combined with screening approaches to pick out the coding sequence that expresses to the highest level. Whilst the randomized libraries are simple and inexpensive to make, the screening steps are usually costly and time-consuming [1].

Translational coupling is a natural phenomenon in bacteria, where initiation of translation is dependent on the successful translation of an upstream sequence, e.g., in the tryptophan operon of *E. coli* between the *trpB* and *trpA* genes [2] or in the tightly controlled stoichiometric expression of genes in the ATP operon

encoding the subunits of the ATP synthase complex [3]. This mechanism has been exploited to synthetically connect the translation of a target gene to a reporter gene without creating protein fusions [4]. In this elegant study, Mendez-Perez and coworkers designed synthetic translational-coupling devices using the knowledge that mRNA secondary structure can mask a Shine-Dalgarno (SD) site, but may re-fold into a less inhibitory structure by the action of an upstream translating ribosome [4]. Applying similar design principles, we created an extended toolbox of sequences that couple with different efficiency and to different antibiotic resistance markers (unpublished). Those reporters, when coupled to the production of a protein of interest, offer specific growth advantages to the host organism and therefore represent a highly attractive alternative to, e.g., fluorescent proteins, enabling screening of very large libraries [5–7]. The extended toolbox of coupling devices enables the selection of gene expression variants from virtually no expression to g/L industrial scale (unpublished).

In this chapter we present our protocols for plasmid construction and selection of high-expressing clones using these translationally coupled antibiotic resistance markers. An overview of the method is presented in Fig. 1. In the first 3 days the selection plasmid is constructed and isolated using uracil excision DNA assembly (Fig. 1a).

The plasmid will contain the coding sequence to be expressed (termed expression variant) fused to a region encoding the translational-coupling device (depicted as an mRNA hairpin) and finally, a region encoding for the antibiotic resistance marker (shown is the coding region for β -lactamase). In the following 4 days a plasmid library is constructed and each plasmid variant is tested by antibiotic sensitivity (Fig. 1b). The library is constructed in a simple PCR reaction using degenerate primers that randomize either the Translation Initiation Region (TIR) or the coding sequence [8, 9]. The entire library is isolated and transformed into a standard *E. coli* expression strain and gene expression is induced in liquid cultures. Bacteria are plated on LB agar plates containing different concentrations of the selectable antibiotic and high-expressing clones can be selected by their ability to withstand the antibiotics. When we applied the protocol to a plasmid encoding a secreted single-chain antibody fragment we were able to select a number of gene variants that expressed to higher levels than the wild type coding sequence (Fig. 1c). These gene variants differed only in the TIR [8]. Shown are Western blots decorated with the corresponding antibody anti-sera (top panel) and a loading control anti-sera (bottom panel). The protocol enables the selection of the high-expressing clones in 1 week starting from initial cloning and using an absolute minimum of resources and only standard molecular biology equipment.

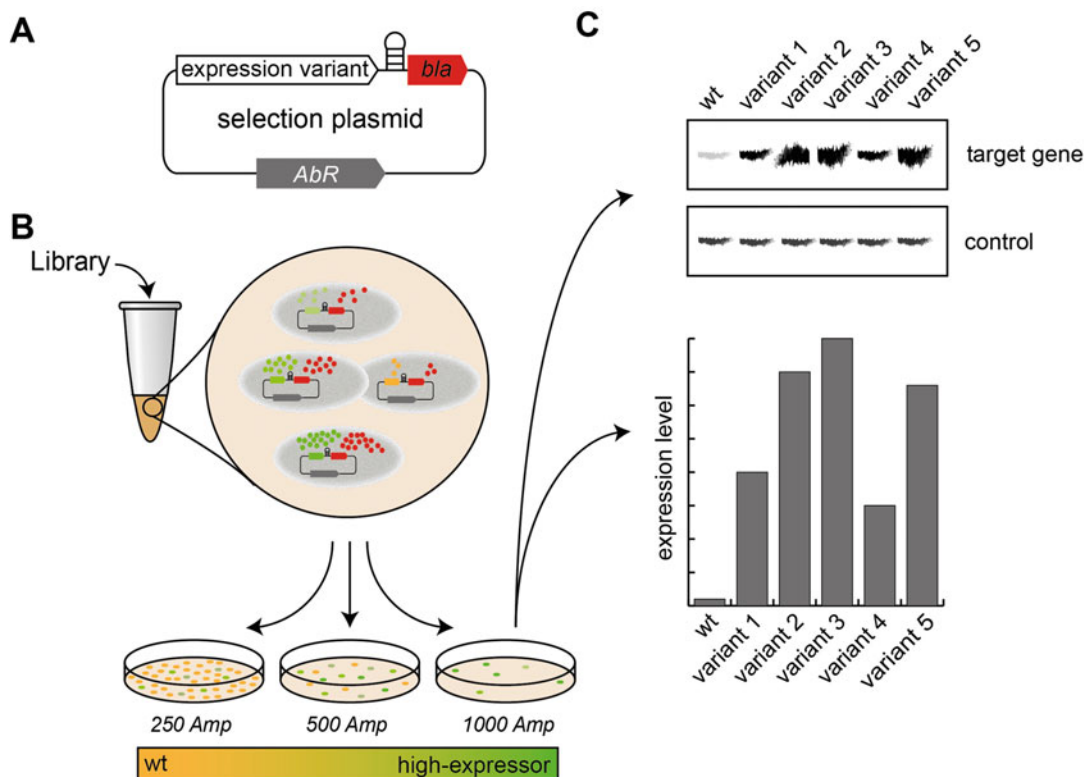


Fig. 1 Schematic illustration of protein production optimization with the translationally coupled antibiotic selection system. (a) Selection plasmid construct with beta-lactamase gene (*bla*) encoding the ampicillin selection marker. (b) Selection of high-expressing variants by plating on LB agar plates containing 0.25, 0.5 and 1 mg/mL ampicillin (Amp). (c) Western blot analysis of the selected optimized variants, on 1 mg/mL ampicillin and expression level estimation of the high-expressing variants

2 Materials

2.1 Components for Plasmid Construction

1. A set of oligonucleotides to integrate the translational coupling element and the antibiotic resistance gene into the desired expression plasmid (*see Note 1* for design).
2. *PfuX7* DNA polymerase with 10× reaction buffer (*see Note 2*).
3. dNTP mix (10 mM each), sterile H₂O and MgCl₂ (final concentration 50 mM) (*see Note 3*).
4. 0.2 mL PCR tubes (VWR).
5. Thermocycler (BioRad).
6. *DpnI* restriction enzyme (New England Biolabs) (*see Note 4*).
7. 1× TAE buffer. Prepare from 50× stock solution (Life Technologies) with H₂O. Can be stored at room temperature.
8. SeaKem[®] LE Agarose (Lonza).

9. RedSafe Nucleic Acid Staining Solution (iNtRON Biotechnology).
10. NucleoSpin[®] Gel and PCR Clean-up Kit (Macherey-Nagel) or any other PCR Clean-up Kit.
11. NanoDrop 2000 UV-Vis Spectrophotometer (Thermo Scientific).
12. 100 ng of each PCR fragment, USER enzyme (New England Biolabs) and provided reaction buffer (*see Note 5*).

2.2 Components for Transformation and Plasmid Propagation

1. *Escherichia coli* strain NEB5 α for cloning. Chemically competent cells of NEB5 α are obtained as described elsewhere [10].
2. Bacteria are cultivated in Luria-Bertani broth (20 g/L in H₂O) (Sigma Aldrich) and plated on Luria-Bertani agar (35 g/L in H₂O) (Sigma Aldrich).
3. Antibiotics: 100 μ g/mL ampicillin, 50 μ g/mL kanamycin, 34 μ g/mL chloramphenicol, 50 μ g/mL spectinomycin and 10 μ g/mL gentamycin (*see Note 6*).
4. 1.5 mL microfuge tubes (Eppendorf).
5. Thermomixer (Eppendorf).
6. 50 mL reaction tubes (Sarstedt).
7. Shaking incubator at 37 °C for 50 mL reaction tubes.
8. Plate incubator at 37 °C.
9. QIAprep Spin Miniprep Kit (Qiagen) or any other plasmid DNA purification Kit.

2.3 Components for Expression and Selection

1. For library expression and selection we use the chemically competent *E. coli* strain BL21(DE3) pLysS (Novagen) (*see Note 7*).
2. Bacteria are cultivated in Luria-Bertani broth (20 g/L in H₂O) (Sigma Aldrich) and plated on Luria-Bertani agar (35 g/L in H₂O) (Sigma Aldrich).
3. 1 M Isopropyl β -D-1-thiogalactopyranoside (IPTG) stock solution in H₂O (*see Note 8*).
4. Antibiotics for selection: ampicillin, kanamycin, chloramphenicol or spectinomycin (*see Note 9*).
5. 50 mL reaction tubes (Sarstedt).
6. Shaking incubator at 37 °C for 50 mL reaction tubes.
7. Plate incubator at 37 °C.

3 Methods

3.1 Primer Design and Selection Set-Up

1. Before constructing the selection plasmid, the expression level of the target gene needs to be evaluated carefully. We have developed a set of different translational coupling devices that vary in their coupling efficiency (Fig. 2).

For low expressed target genes we recommend the strong coupling device. For genes already expressing to decent levels, a weaker coupling device is advisable. This way the use of very high antibiotic concentrations for selection is avoided and the dynamic range of the selection system is better exploited (*see Note 10*). The coupling device is introduced with PCR oligonucleotides - the major part with a “forward” oligonucleotide for amplification of the antibiotic resistance gene (here denoted AbR gene fwd). The reverse oligonucleotide for amplifying the target gene introduces the second part of the coupling device (here denoted target gene rev.) (Fig. 3).

2. The target gene can either be amplified together with the backbone or can be cloned into a new backbone. When amplified with the backbone primer set 1 is not needed (Fig. 3). In case that primer set 1 is needed, the design will be similar to primer set 3. The reverse primer for amplifying the target gene introduces parts of the coupling device (target gene rev.) and the only necessary modification to the gene of interest ensures that TGA is used as a stop codon. Moreover, the reverse primer for

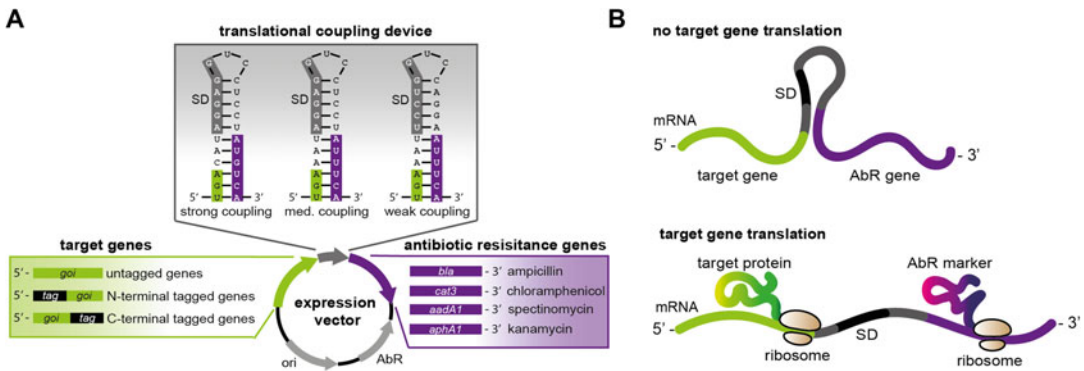


Fig. 2 Different modules can be combined for optimal antibiotic selection via translational coupling. (a) The antibiotic selection system optimizes the expression of untagged, N-terminal tagged, or C-terminal tagged genes. The target gene (green box) is combined with one out of three translational coupling devices (grey box). The three different hairpins vary in their coupling efficiency due to altered start codons (shaded purple regions) and Shine-Dalgarno (SD) sequences (shaded grey regions). Four different antibiotic resistance genes (AbR, purple box) can be used as reporters. (b) Illustration of the principle of translational coupling; if the mRNA from the target gene (green) is not translated, the ribosome-binding site (black) for the reporter gene (purple, AbR gene) is not accessible. If the target gene is translated, the ribosome will melt the hairpin structure and enable a ribosome to bind to the SD sequence of the reporter gene mRNA and start translation

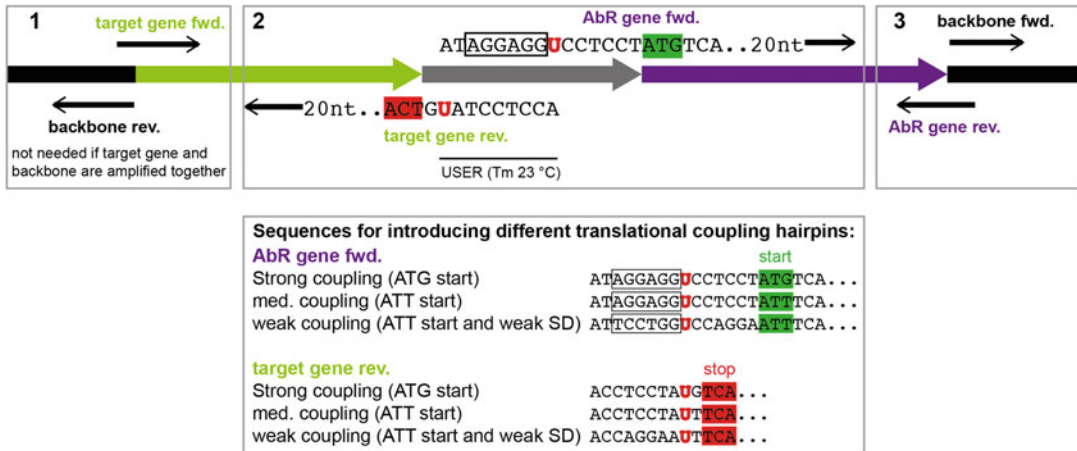


Fig. 3 Illustration of the design of oligonucleotides that introduce a translational coupling device and an antibiotic resistance gene for selection. Primer pairs denoted in boxes 1, 2, and 3 have complementary 5' ends as exemplified with sequence detail in box 2

amplifying the target gene should ideally anneal with a melting temperature of ca. 60 °C to the sequence upstream from the stop codon. Typically this takes roughly 20 nucleotides for the anneal part resulting in a total oligonucleotide size of ca. 36.

- The major part of the translational coupling device is introduced with the forward primer for amplification of the antibiotic resistance gene (AbR gene fwd.). Besides encoding the device, the primer should anneal with a melting temperature of ca. 60 °C to sequence downstream from the start codon. This will typically result in oligonucleotide size of ca. 40 nucleotides. Note that the first amino acid after the start codon is still part of the coupling device and will always be TCA (encoding for serine). The ampicillin and kanamycin resistance genes, *bla* and *aphA1*, already encode serine as the second amino acid. For the spectinomycin and chloramphenicol resistance genes, *aadA1* and *cat*, the addition of serine in this position does not influence functionality (*see Note 11*).
- Following the PCR, cohesive ends with a melting temperature of ca. 23 °C are formed by uracil excision (*see Note 12*) and the resulting nicked circular DNA can be transformed into chemically competent *E. coli*. Details about uracil excision DNA assembly can be found elsewhere [11].

3.2 Plasmid Construction and Propagation

- Mix all reagents for a 50 µL PCR reaction in a 0.2 mL PCR tube: 5 µL of 10× reaction buffer, 5 µL of forward primer, 5 µL of reverse primer, 2 µL dNTP mix (40 mM), 1.2 µL 2 M MgCl₂, 1 µL PfuX7 DNA polymerase, 100 ng of template DNA, fill up to 50 µL with sterile H₂O.

2. Amplify the different fragments in a Thermocycler using a Touchdown-PCR program (95 °C for 5 min, 10 cycles of 95 °C for 45 s, 55–65 °C for 45 s (decrement 1 °C in each cycle), 72 °C for 5 min, followed by 20 cycles of 95 °C for 45 s, 55 °C for 45 s, 72 °C for 5 min and one final elongation step for 8 min at 72 °C (*see Note 13*).
3. Prepare a 1% agarose gel by heating up 1 g of agarose in 100 mL of 1× TAE buffer. Add DNA stain. Analyze 5 µL of the PCR product by agarose gel electrophoresis (e.g., 90 V, 30 min).
4. Add 2 µL *DpnI* to the reaction mix to digest the original plasmid at 37 °C for at least 1 h. Inactivate *DpnI* for 5 min at 80 °C.
5. Purify the PCR fragments using a PCR Clean-up Kit, following the manufacturer's instructions.
6. Mix 100 ng of each fragment. Add 1 µL of USER enzyme and 1 µL of buffer. Fill up with sterile H₂O to a final volume of 10 µL.
7. Start the USER reaction in a Thermocycler: 37 °C for 15 min, 23 °C for 15 min, 10 °C for 15 min (*see Note 14*).
8. Mix the whole USER reaction with 100 µL of chemically competent *E. coli* cells. Use a cloning strain, such as NEB5α.
9. Incubate on ice for 30 min.
10. Heat shock for 1 min at 42 °C.
11. Incubate on ice for 3 min.
12. Add 0.9 mL of LB broth without antibiotics for recovery and incubate at 37 °C with shaking for 1 h.
13. Plate 0.1 mL (10%) of the transformation mix on a LB agar plate containing the appropriate antibiotics.
14. Harvest the rest of the transformation mix by centrifugation at 4000 × *g* in a tabletop centrifuge and plate on an LB agar plate containing the appropriate antibiotics.
15. Incubate for 16 h at 37 °C.
16. Inoculate 5 mL LB media containing the appropriate antibiotics (in a 50 mL reaction tube) with a single colony from the transformation plates and incubate for 16 h at 37 °C with shaking.
17. Harvest the cells at 6500 × *g* and purify the plasmid using a plasmid DNA purification Kit (e.g., QIAprep Spin Miniprep Kit), follow the manufacturer's instructions.
18. Prepare an expression library (*see Note 15*).

3.3 Expression and Selection

1. Transform the library and the wild type plasmid into an expression strain, e.g., *E. coli* BL21(DE3) pLysS. To do so, follow **steps 8–12** in Subheading **3.2**. Transform ca. 5 μ L of library into 50 μ L of commercially competent cells, and 1 μ L of wild type plasmid into 15 μ L of commercially competent cells. Heat shock for 45 s at 42 °C.
2. After 1 h of recovery transfer the transformation mix to a 50 mL reaction tube containing 5 mL of LB media with appropriate antibiotics and incubate for 16 h at 37 °C with shaking.
3. Prepare plates for selection. Use LB agar containing the inducing agent, we use IPTG in a final concentration of 1 mM, and different concentrations of the selective antibiotic (*see Note 16*). We recommend preparing selection plates with at least five different antibiotic concentrations. Remember to prepare two plates of each concentration, one for the library and one for the wild type.
4. After 16 h, start an expression culture by inoculating 5 mL fresh LB medium containing appropriate antibiotics with 100 μ L of overnight culture (1:50). Incubate at 37 °C with shaking.
5. Grow the culture to an OD₆₀₀ of 0.3–0.5 (ca. 2 h) and induce expression with IPTG (final concentration 1 mM) or any other agent needed to induce expression. Incubate at 37 °C with shaking.
6. After 2 h of expression, plate 0.2 ODU (ca. 100 μ L) on each selection plate. Incubate for up to 40 h at 37 °C. On antibiotic concentrations where no growth can be seen for the wild type, high expressing variants can be selected.
7. Select variants and check with sequencing (*see Note 17*).
8. Re-transform the selected variants to confirm the selection of a highly translated variant. Follow **steps 1–5** (skip **step 3**) and check for expression levels, e.g., by Western blot analysis.

4 Notes

1. The forward primer of the target gene and the reverse primer of the antibiotic resistance gene form the translational coupling device. Oligonucleotides contain an incorporated uracil for USER cloning.
2. Alternatively Phusion U Hot Start DNA Polymerase (Thermo Scientific) can be used.
3. If Phusion U Hot Start DNA Polymerase is used, 1.5 μ L of 1 M DMSO is added to a 50 μ L reaction.

4. 2 μ L FastDigest *DpnI* (Thermo Scientific) were added directly to a 50 μ L PCR mix after PCR reaction, since *DpnI* works effectively in the reaction buffers used for PCR.
5. We recommend a ligation step after the USER reaction, if electro-competent cells are used. In this case, we recommend using T4 ligase buffer for the USER reaction.
6. Antibiotic concentrations indicated are those that should be used for the backbone antibiotic resistance. Make sure that the backbone resistance and the resistance used for selection are different.
7. Be aware that when using a plasmid with chloramphenicol resistance, either in the backbone or as selection module, BL21(DE3) without pLysS needs to be used, as the pLysS plasmid (and derivatives) confer resistance to chloramphenicol.
8. The type of inducing agent depends on the promoter used. We are using a (DE3) T7 promoter and therefore typically use IPTG for induction.
9. Concentrations depend on the target gene expression level and the chosen coupling efficiency.
10. The coupling device can be changed easily in a one-PCR step followed by a one-fragment USER reaction.
11. Note that the chosen resistance gene for selection should not be present as a backbone resistance gene on any other plasmid transformed. We obtained the best results when using the ampicillin resistance gene for selection.
12. We recommend choosing a USER cloning overlap in the backbone sequence, as it facilitates the exchange of the antibiotic resistance gene, if necessary.
13. Note that the extension time depends on the size of the PCR product and the processivity of the DNA polymerase that is used.
14. When using electro-competent cells, add 1.5 μ L of T4 ligase and incubate for 15 min at room temperature prior to transformation.
15. Library construction is not part of this protocol. Sequence variation may be sampled throughout the gene coding sequence or expression vector. We construct libraries based on the protocol “Codon optimizing for increased membrane protein production: A minimalist approach” [9]. Additionally, we expand the library by also randomizing the six nucleotides upstream of the start codon [8].
16. Make sure that the LB agar has cooled down to at 60 °C before adding the inducing agent and antibiotics.

17. Sequencing after selection is of utmost importance. It can occur that the cell starts mutating in ways to favor the expression and usage of the antibiotic resistance gene and a false positive is selected.

References

1. Rosano GL, Ceccarelli EA (2014) Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front Microbiol* 5:1–17
2. Aksoy S, Squires CL, Squires C (1984) Translational coupling of the *trpB* and *trpA* genes in the *Escherichia coli* tryptophan operon. *J Bacteriol* 157:363–367
3. Rex G, Surin B, Besse G et al (1994) The mechanism of translational coupling in *Escherichia coli*. Higher order structure in the *atpHA* mRNA acts as a conformational switch regulating the access of de novo initiating ribosomes. *J Biol Chem* 269:18118–18127
4. Mendez-Perez D, Gunasekaran S, Orler VJ et al (2012) A translation-coupling DNA cassette for monitoring protein translation in *Escherichia coli*. *Metab Eng* 14:298–305
5. Massey-Gendel E, Zhao A, Boulting G et al (2009) Genetic selection system for improving recombinant membrane protein expression in *E. coli*. *Protein Sci* 18:372–383
6. Gul N, Linares DM, Ho FY et al (2014) Evolved *Escherichia coli* strains for amplified, functional expression of membrane proteins. *J Mol Biol* 426:136–149
7. Tan R, Jiang X, Jackson A et al (2003) *E coli* selection of human genes encoding secreted and membrane proteins based on cDNA fusions to a leaderless β -lactamase reporter. *Genome Res* 13:1938–1943
8. Mirzadeh K, Martínez V, Toddo S et al (2015) Enhanced protein production in *Escherichia coli* by optimization of cloning scars at the vector–coding sequence junction. *ACS Synth Biol* 4:959–965
9. Mirzadeh K, Toddo S, Nørholm MHH et al (2016) Codon optimizing for increased membrane protein production: a minimalist approach. In: *Heterologous expression of membrane proteins, Methods and protocols, methods in molecular biology*. Springer Science+Business Media, New York, pp 53–61
10. Inoue H, Nojima H, Okayama H (1990) High efficiency transformation of *Escherichia coli* with plasmids. *Gene* 96:23–28
11. Cavaleiro AM, Kim SH, Seppälä S et al (2015) Accurate DNA assembly and genome engineering with optimized uracil excision cloning. *ACS Synth Biol* 4:1042–1046

Chapter 17

Design, Engineering, and Characterization of Prokaryotic Ligand-Binding Transcriptional Activators as Biosensors in Yeast

Francesca Ambri, Tim Snoek, Mette L. Skjoedt, Michael K. Jensen, and Jay D. Keasling

Abstract

In cell factory development, screening procedures, often relying on low-throughput analytical methods, are lagging far behind diversity generation methods. This renders the identification and selection of the best cell factory designs tiresome and costly, conclusively hindering the manufacturing process. In the yeast *Saccharomyces cerevisiae*, implementation of allosterically regulated transcription factors from prokaryotes as metabolite biosensors has proven a valuable strategy to alleviate this screening bottleneck. Here, we present a protocol to select and incorporate prokaryotic transcriptional activators as metabolite biosensors in *S. cerevisiae*. As an example, we outline the engineering and characterization of the LysR-type transcriptional regulator (LTTR) family member BenM from *Acetivobacter* sp. ADPI for monitoring accumulation of *cis,cis*-muconic acid, a bioplast precursor, in yeast by means of flow cytometry.

Key words Biosensor, Transcription factor, Cell factory, Synthetic biology, Screening, Yeast

1 Introduction

In the last two decades, metabolic engineering has proven itself as a key technology for the manufacturing of valuable molecules from renewable feedstocks [1–3]. The constant development of synthetic biology tools for cloning, library construction, and genome engineering, coupled with the sharp decrease in DNA synthesis costs, has greatly aided the process of building more efficient microbial cell factories [1–3]. Adversely, when aiming to engineer a high-performing cell factory, most heterologous compounds do not give a clear phenotype, rendering the screening procedure of a population of cell factory designs through conventional spectrophotometry-based analytics methods slow and often costly.

Francesca Ambri and Tim Snoek contributed equally to this work.

Michael Krogh Jensen and Jay D. Keasling (eds.), *Synthetic Metabolic Pathways: Methods and Protocols*, Methods in Molecular Biology, vol. 1671, https://doi.org/10.1007/978-1-4939-7295-1_17, © Springer Science+Business Media, LLC 2018

The fact that screening and selection methods lag far behind the techniques that allow the generation of large libraries ultimately hampers the identification of cell factory designs with optimal titers, rates, and yields.

Genetically encoded biosensors convey a huge potential to overcome this screening bottleneck. In general, biosensors are able to sense input, like extra- or intracellular metabolite perturbations, and subsequently actuate an adequate output akin to logic gates in electrical circuits. Natural biosensors include molecular gating components like RNA aptamers and allosterically regulated transcription factors, which can regulate transcription of target genes in the presence of an adequate input [1–6]. In the case of ligand-binding allosterically regulated transcription factors, placing a reporter gene or non-native selection gene under the control of the ligand-binding transcription factor offers a sensitive and specific synthetic system that can couple input (intracellular ligand concentration) to a high-throughput screenable output (e.g. fluorescence or antibiotic resistance) [1–7].

Prokaryotes harbor an enormous reservoir of ligand-binding transcriptional regulators, which can be applied as biosensors [8]. These transcriptional regulators can be grouped into transcriptional repressors and transcriptional activators (Fig. 1). Transcriptional repressors have abundantly been applied as biosensors both in prokaryotes and eukaryotes [1–7]. Perhaps the most famous example is the TetR system. Specifically, when tetracyclin (tc) is absent, the transcription factor TetR represses the expression of a membrane protein (TetA) through binding to an operator in the *tetA* promoter (TetO), thereby preventing transcription. In contrast, the presence of tc induces a conformational change of TetR rendering it unable to bind TetO and thereby relieving the repression of *tetA* expression, ultimately forcing tc out of the cell [9]. Because of the efficiency of TetR to sense sub-inhibitory tc concentrations ($K_a = 10^{10} \text{ M}^{-1}$) and its high specificity for TetO, the prokaryote TetR-TetO system has been engineered and transplanted into eukaryote hosts for monitoring and regulation purposes [10–12]. Other examples of prokaryotic repressors that have been successfully transplanted as biosensors into *Saccharomyces cerevisiae* include FapR, for the detection of malonyl-CoA, and XylR, detecting xylose [13–15].

Until recently, transcriptional activators had not been reported as metabolite biosensors in a eukaryote, limiting the reservoir of transcriptional regulators that can be tapped from for biosensor development. The different mode of action for activators compared to repressors, as well as the inherent need for more extensive engineering of the biosensor in order obtain a relevant input to output relationship (Fig. 1), are possible explanations for why prokaryotic transcriptional activators had not been transplanted as biosensors into eukaryotes yet. Indeed, in our recent study, both

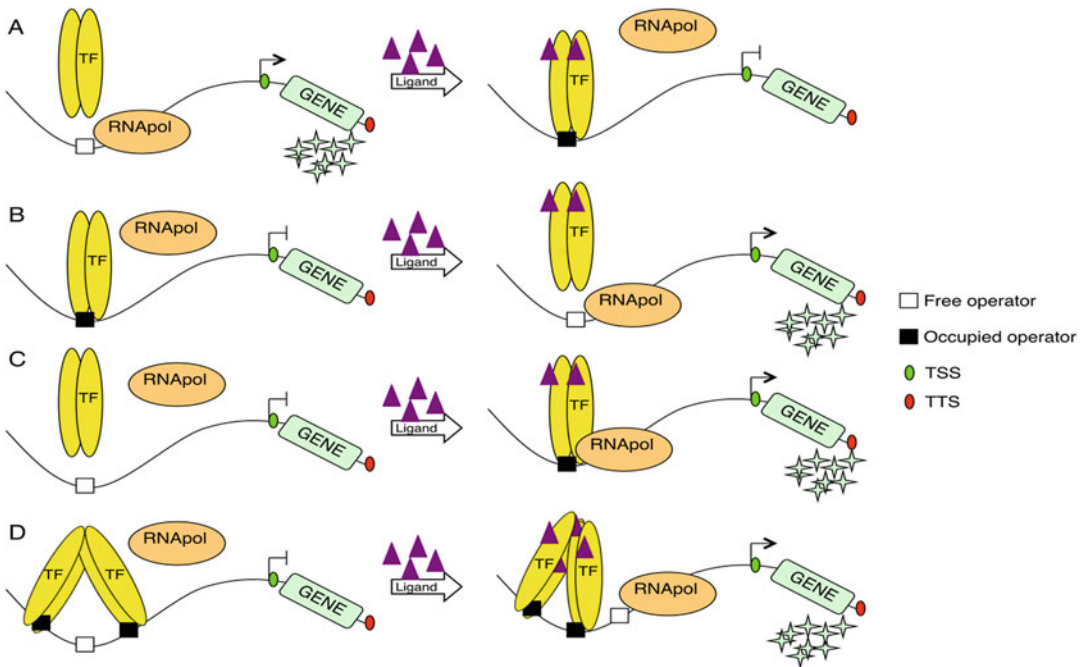


Fig. 1 Mechanisms for gene expression regulation. (a) Repressor-based repression: in the presence of the ligand, the transcription factor (TF) binds to the operator site within the promoter region and physically hinders the RNA polymerase activity; (b) Activator-based regulation: in the presence of the ligand, the TF acts as a recruiter for RNA polymerase when bound to the operator site; (c) Repressor-based activation: the repressor is constitutively bound to the operator site and the presence of the ligand triggers a conformational change in the TF impairing the binding, consequently relieving the repression; (d) BenM-based gene regulation in *Acinetobacter* sp. ADP1: CCM, a degradation compound from aromatic acid catabolism, triggers a conformational change in BenM that results in a shift of the sites being bound, facilitating access of RNA polymerase and activate transcription of the downstream operon (see **Note 1** for more details)

tuning of the expression level of transcriptional activators and extensive engineering of reporter promoters turned out to be important for establishing small-molecule biosensors based on prokaryotic transcriptional activators [16].

In this protocol we present the (1) selection criteria, (2) design, (3) engineering, and (4) application of a biosensor for *cis,cis*-muconic acid (CCM) in the budding yeast *S. cerevisiae* as we previously reported [16]. CCM is an important bioplastics precursor, the production of which from renewable carbon sources has been engineered in yeast [17]. In the bacterium *Acinetobacter* sp. ADP1, BenM is a LysR-type transcriptional regulator (LTTR) that upon binding of CCM undergoes a conformational change resulting in transcriptional activation of genes involved with catabolism of aromatic compounds (Fig. 1) [18]. BenM is a well-studied gene and protein; with annotated gene sequence, crystal structure, and operator sequence [19]. In our original study, we optimized various parameters to establish BenM as a CCM biosensor in yeast,

and proved the validity of the biosensor design for other biosensors based on LTTR-type of regulators. In this protocol we will focus on the engineering of the optimal CCM biosensor design and how to characterize its performance.

2 Materials

2.1 Strains, Media and Reagents

1. Bacterial strain: *Escherichia coli* DH5 α is used as a host for cloning and plasmid propagation.
2. Yeast strain: *S. cerevisiae* CEN.PK113-5A (MAT α , *trp1 his3 Δ 1 leu2-3/112 MAL2-8^c SUC2*) is used as the basic strain in which the biosensor-reporter will be built in.
3. Media: For *E. coli*; Luria-Bertani (LB) medium with ampicillin (10 g tryptone, 5 g yeast extract, 10 g NaCl, deionized water up to 1 L, autoclave, add 100 μ g/mL ampicillin). For *S. cerevisiae*; Synthetic Complete (SC) medium (6.7 g yeast nitrogen base without amino acids, appropriate amount of drop-out medium supplement (Sigma-Aldrich), deionized water up to approximately 880 mL, adjust pH to 5.6, deionized water up to 900 mL, (20 g agar in case of plates), autoclave, 100 mL 20% (w/v) glucose); mineral medium with tryptophan (per L: 7.5 g (NH₄)₂SO₄, 14.4 g KH₂PO₄, 0.5 g MgSO₄·7H₂O, 20 g glucose, 2 mL trace metals solution, 1 mL vitamin solution, 4 mL tryptophan solution (5 g/L)); and yeast extract peptone dextrose (YPD) complete medium (10 g BactoYeast extract, 20 g BactoPeptone, 20 g Dextrose, (20 g agar in case of plates), deionized water up to approximately 1000 mL, autoclave); pre-mineral medium (75 mL ammonium solution (NH₄)₂SO₄ (100 g/L), 120 mL phosphate solution KH₂PO₄ (120 g/L), 10 mL magnesium solution MgSO₄·7H₂O (50 g/L), 100 mL 20% (w/v) glucose, 2 mL trace metals, 1 mL vitamins, 4 mL 250 \times tryptophan stock solution (5 g/L) deionized water up to 900 mL, mix using stirring rod, filter sterile, store at 4 °C).

2.2 Molecular Biology

1. DNA polymerases: High-fidelity Phusion U Hot Start DNA Polymerase (Thermo Fisher Scientific, Inc.), 2xOneTaq[®] Master Mix DNA Polymerase (New England Biolabs).
2. Synthetic genes: Commercially synthesized (Integrated DNA Technologies, Inc.). Genes are codon-optimized for expression in yeast using manufacturer's software (Table 1).
3. Oligonucleotides: Commercially synthesized (Integrated DNA Technologies, Inc.) (Tables 2, 3, 4).

Table 1
Synthetic DNA fragments

| |
|--|
| BenM coding sequence codon-optimized for <i>S. cerevisiae</i> : |
| <pre> 1 ATGGAATTGA GACTTGTGAG ATACTTCGTT GCCGTTGTTG AAGAACAATC TTTTACAAAG 61 GCTGCCGACA AGTTGTGTAT TGCTCAACCA CCATTATCCA GACAAATCCA AAAC TTGGAA 121 GAAGAATTGG GTATCCAATT ATTGGAAAGA GGTTCAGAC CAGTTAAGAC TACTCCAGAA 181 GGTCATTTCT TTTACCAATA CGCCATCAAG TTGTTGTCCA ACGTTGATCA AATGGTCAGT 241 ATGACCAAGA GAATTGCCTC TGTTGAAAAG ACCATTAGAA TCGGTTTTGT TGGTTCCTTG 301 TTGTTGCGTT TGTTGCCAAG AATTATCCAC TTGTACAGAC AAGCTCATCC AAAC TTGAGA 361 ATCGAATTAT ACGAAATGGG TACTAAGGCT CAAACCGAAG CTTTGAAAGA AGGTAGAATT 421 GACGCTGGTT TTGGTAGATT GAAGATTTCT GATCCAGCCA TCAAGAGAAC CTGTTGAGA 481 AACGAAAGAT TGATGGTTGC TGTTTCATGCT TCCCATCCAT TGAATCAAAT GAAGGATAAG 541 GGTGTTCACT TGAACGATTT GATCGACGAA AAGATCTTGT TGTACCCATC TTCTCCAAAG 601 CCAAATCTCT CTACTCATGT TATGAACATC TTCTCTGACC ATGGTTTTGA ACCTACCAAG 661 ATTAACGAAG TTAGAGAAGT CCAATTGGCC TTGGGTTTTG TTGCTGCTGG TGAAGGTATT 721 TCATGGTTC CAGCTTCTAC CCAATCCATT CAATTATTCA ACTTGTCTTA CGTCCATTA 781 TTAGATCCAG ATGCTATTAC CCCAATCTAC ATTGCTGTTA GAAACATGGA AGAATCCACC 841 TACATCTACT CATTATACGA AACCATCAGA CAAATCTACG CCTACGAAGG TTTTACTGAA 901 CCACCAAATT GGTA </pre> |
| Sequence of 209bp_CYC1p_BenO_T1: the minimal <i>CYC1</i> promoter with <i>BenO</i> (underlined) inserted 6 bp upstream of TATA-1 β (indicated in bold): |
| <pre> 1 CCAGGCAACT TTAGTGCTGA CACATAATAC TCCATAGGTA TTTTATTATA CAAATAATGT 61 GTTTGAAC TT ATTA AACAT TCTTTTAAAGG TATAAACAA C AGGCATATAT ATATGTGTGC 121 GACGACACAT GATCATATGG CATGCATGTG CTCTGTATGT ATATAAACT CTTGTTTTCT 181 TCTTTCTCT AAATATCTT TCCTTATACA TTAGGACCTT TGCAGCATAA ATTACTATAC 241 TTCTATAGAC ACACAAACAC AAATACACAC ACTAAATTAA TA </pre> |

4. Plasmids: The EasyClone plasmids used in this protocol are from Jensen et al. [20]. All constructed plasmids are sequence-verified by Sanger sequencing.
5. Gel purification: Amplified genes, promoters and digested vectors are gel-purified using NucleoSpin[®] Gel and PCR Clean-up kit (Machery-Nagel).
6. Restriction enzymes: FastDigest[®] SfaI and FastDigest[®] NotI with the corresponding FastDigest[®] buffer; Nb.BsmI and corresponding buffer 3.1 (New England Biolabs).
7. USER cloning: USER[™] enzyme (New England Biolabs), 5 \times Phusion HF Buffer (Life Technologies).

2.3 Reagents

1. 0.02 M NaOH is used for yeast cell lysis.
2. 1% Agarose is used for routine analysis of nucleic acids by gel electrophoresis.

Table 2
Primers for USER cloning of genes and promoters

| Description primer | Sequence 5'-3' |
|---|--|
| Forward primer for USER cloning of REV1p | CGTGCGA U TTCTTAGGCACAACA TATTTATAAAAAGAAG |
| Reverse primer for USER cloning of REV1p | ATGACAGA U CGCTGGATATGCCTAGAA ATGC |
| Forward primer for USER cloning of BenM (Kozak, start codon) | ATCTGTCAU AAAA CAATGGAATTGAGACAC |
| Reverse primer for USER cloning of BenM | CACGCGA U TACCAATTTGGTGGTTCAG |
| Forward primer for USER cloning of 209bp_CYC1p_BenO_T1 | CGTGCGA U CCAGGCCAACTTTAGTGCTGACAC |
| Reverse primer for USER cloning of 209bp_CYC1p_BenO_T1 | ATGACAGA U TATTAATTTAGTGTGTGTATTTGTGTTTGTG |
| Forward primer for USER cloning of yeGFP (Kozak, start codon) | ATCTGTCAU AAAA CAATGTCTAAAGGTG |
| Reverse primer for USER cloning of yeGFP | CACGCGA U TATTTGTACAATTCATCCA |

Uracil-containing overhangs are indicated in bold

Table 3
Universal primers binding common EasyClone backbone vectors are used to confirm the correct assembling of the expression cassette into the integrative vector backbone

| Primer ID | Sequence 5'-3' | Description |
|-----------|---------------------------|--------------------------|
| 224 | GAAATTCGCTTATTTAGAAAGTGTC | Universal forward primer |
| 225 | CTCCTTCCTTTTCGGTTAGAG | Universal reverse primer |

3. Phosphate buffered saline (PBS) is used for cell dilution prior to flow cytometry.
4. *Cis,cis*-muconic acid (Sigma-Aldrich).
5. Trace metals solution (per L: 4.5 g CaCl₂·2H₂O, 4.5 g ZnSO₄·7H₂O, 3 g FeSO₄·7H₂O, 1 g H₃BO₃, 1 g MnCl₂·4H₂O, 0.4 g Na₂MoO₄·2H₂O, 0.3 g CoCl₂·6H₂O, 0.1 g CuSO₄·5H₂O, 0.1 g KI, 15 g EDTA. Add the salts (without EDTA) one by one to 900 mL deionized water and dissolve them, while keeping the pH at 6. Then, gently heat the solution and add EDTA. Finally, adjust pH to 4, bring final volume to 1 L, autoclave and store at 4 °C).

Table 4

Primers 2220 and 2221 are universal primers annealing to any EasyClone vector of choice. The other primers correspond to an insertion into a particular genomic integration site. Either one or both primer pairs can be used for verification of the particular correct integration

| Site | Primer ID | Sequence 5'-3' | Description | Fragment size (bp) |
|-------|-----------|------------------------------|-------------|--------------------|
| X-3 | 2220 | CCTGCAGGACTAGTGCTGAG | X-3 DOWN | 667 |
| | 904 | CCGTGCAATACCAAATCG | | |
| | 2221 | GTTGACACTTCTAAATAAGCGAATTTTC | X-3 UP | 1059 |
| | 903 | TGACGAATCGTTAGGCACAG | | |
| XII-4 | 2220 | CCTGCAGGACTAGTGCTGAG | XII-4 DOWN | 667 |
| | 898 | CGTGAAATCTCTTTGCGGTAG | | |
| | 2221 | GTTGACACTTCTAAATAAGCGAATTTTC | XII-4 UP | 828 |
| | 897 | GAAC TGACGT CGAAGGCTCT | | |

- Vitamin solution (per L: 50 mg biotin, 200 mg p-aminobenzoic acid, 1 g nicotinic acid, 1 g Ca-pantothenate, 1 g pyridoxine-HCl, 1 g thiamine-HCl, 25 g myo-inositol. Dissolve biotin in 20 mL 0.1 M NaOH, and then add 900 mL deionized water. Adjust pH to 6.5 and add the remaining vitamins. Readjust pH to 6.5 just before and after adding m-inositol. Adjust the final volume to 1 L. Filter-sterilize and store at 4 °C).
- Yeast transformation mix: 50% w/v Polyethylene glycol solution MW 3350 (PEG3350), 1 M Lithium acetate pH 7.5 (LiAc), 2 mg/mL Deoxyribonucleic acid (DNA) single stranded from salmon testes (ssDNA) dissolved in sterile TE (10 mM Tris-HCl, 1 mM Na₂EDTA pH 8.0) boil for 5 min and then keep on ice (Sigma-Aldrich).

2.4 Flow Cytometry

- Culturing: Regular 96-well plates (volume: 360 µL per well) are used for pre-culturing, polypropylene deep well plates (volume: 1 mL per well) are used for the subculturing in induction (+ CCM) or control (-CCM) medium prior to flow cytometry analysis.
- Instrument: Becton Dickinson LSR FORTESSA with a blue 488 nm laser.
- Software: BD FACSDIVA™ is used by the BD LSR FORTESSA machine for data acquisition; FlowJo is used for data analysis.

3 Methods

In the following section the step-by-step design, engineering, and characterization of BenM as a CCM biosensor in *S. cerevisiae* is described (*see Note 2*). The procedure starts with identification of

the candidate transcription factor and its corresponding operator sequence and ligand from literature. Based on this information, the protocol describes how to generate a yeast strain that incorporates a sensor-reporter system that basically consists of two constructs: (1) the gene encoding the TF driven by a yeast promoter, and (2) the yeast-enhanced GFP (yeGFP) reporter gene driven by an engineered yeast promoter with the operator sequence for the ligand-binding TF. For the former, it is generally advisable to compare different expression levels (i.e. promoters with various strengths), whereas for the latter, the design of the reporter promoter, in particular the number and positioning of the operator sequence, is important [9, 21]. In the case of transcriptional activator-based biosensors, an ideal scenario is a low basal activity of the reporter (output) in the absence of the ligand (input), combined with a strong dose-response output with increasing concentrations of the target ligand. Indeed, finding this optimal input–output configuration required the generation of a library of strains, in which each reporter promoter variant was combined with various expression levels of the TF. The resulting library covered a wide range of modulated responses to the same inducer enabling a more reliable and sensitive detection of the ligand permitting ad hoc high-throughput screening of strains [16]. For simplicity, this protocol only outlines the design and engineering of the optimal sensor-reporter design as identified in our study [16]. Next, this protocol also describes how to characterize the performance of the biosensor using flow cytometry.

3.1 Selection of Candidate Biosensor

The first step to construct a transcription factor-based biosensor is the identification of a transcription factor, which can bind, or in other indirect ways sense, the chemical of interest.

Consider whether the following criteria are fulfilled:

1. Gene sequence of the ligand-binding TF is known.
2. Operator sequence of the TF is known.
3. The ligand is commercially available, is not toxic to yeast at relevant concentrations, and can be taken up by yeast at a tolerable pH (3.5–7.5).
4. Optional: Consideration of the need for expression of additional genes in yeast that would allow for the TF to function (e.g. cofactors) or for the ligand to be taken up (e.g. transporters) (*see Note 3*).

Predictably the starting point to identify these elements is literature mining for either publications of transcription factor discovery, operator sequence, molecule of interest, motifs' similarity or distinctive characteristics of the system under study. In addition, public databases are available for browsing the genome of many species enabling a direct approach when looking for operator sites (e.g. <http://www.pseudomonas.com/>).

In the case of BenM, both gene sequence and operator sequence are known, CCM is commercially available and its production has been engineered in yeast before [17–19].

3.2 Preparation of Gene and Promoter Fragments for USER Cloning

The next step is the design and ordering of synthetic DNA constructs for both the gene encoding the transcription factor and the reporter promoter, as well as primers that can amplify these parts, which will result in fragments for cloning into the appropriate vectors. In order to robustly characterize biosensor designs, it is advisable to stably integrate both the sensor and the reporter construct into defined genomic loci. In this protocol USER cloning and the EasyClone system are used for creating integrative vectors and inserting the resulting constructs into the genome using homologous recombination [20, 22].

3.2.1 Design of Synthetic DNA Constructs

In order to engineer a genetically encoded biosensor based on a transcription factor, two synthetic DNA constructs have to be synthesized (Table 1):

1. Open reading frame BenM (sensor): copy and paste gene sequence (*see Note 4*) and use the codon optimization for *S. cerevisiae* option in the web tool on the IDT website (<http://eu.idtdna.com/CodonOpt>).
2. Reporter promoter: In the optimal CCM biosensor design identified in our study, the BenM operator sequence is integrated 6 bp upstream of TATA-I β in the truncated 209 bp CYC1 minimal promoter [16]. We will refer to this promoter as 209bp_CYC1p_BenO_T1.

In addition to these two synthetic constructs, two other parts, i.e. the promoter driving expression of the transcription factor, and the reporter gene coding for yeGFP (*see Note 5*), can be obtained from amplification of genomic DNA or commercially available resources (e.g. Addgene plasmid ID 40235), respectively. We will illustrate the example for constructing the optimal CCM biosensor using the weak *REV1* promoter (REV1p) to control the expression of BenM.

3.2.2 Primer Design

Primers with uracil overhangs are used for the amplification of genes and promoters. As illustrated in the article describing the EasyClone system [20], the overhangs vary depending on the chosen combination of promoter-gene and insertion position in the backbone [20]. In this example genes and promoters are always inserted into position 2 (*see Fig. 2*).

3.2.3 PCRs

All the genes and promoters are amplified with primers containing uracil, which requires the use of a DNA polymerase that can read through uracil without hindering its proofreading activity (Fig. 2a).

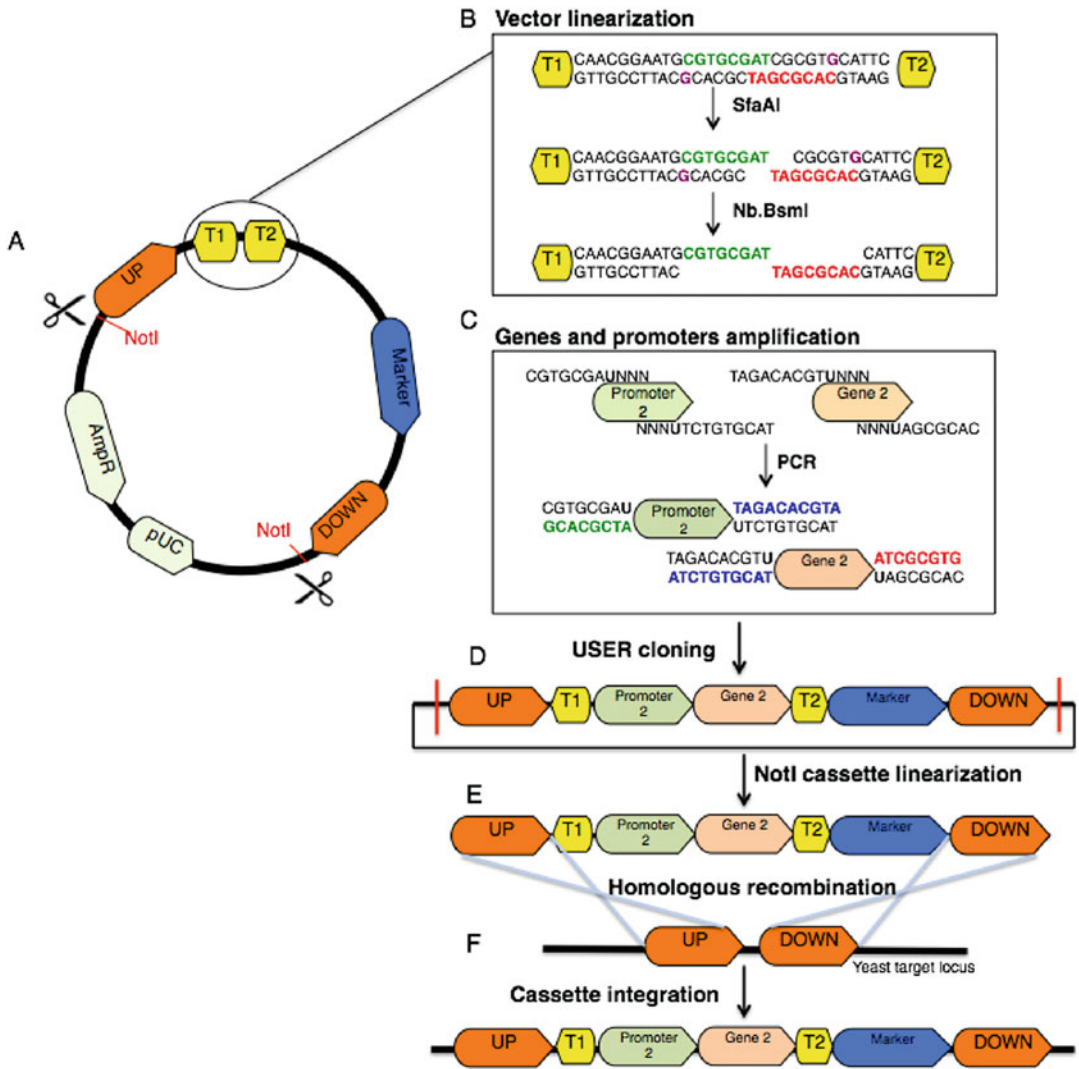


Fig. 2 Cloning genes and promoters into EasyClone vectors: (a) The basic vector backbone includes genetic element for vector replication—pUC—and selection—AmpR—in bacteria; NotI restriction sites flanking yeast homologous regions—UP and DOWN; an auxotrophic selectable marker gene; two yeast terminators—ADH1 and CYC1—T1 and T2 separated. (b, c) A two-step digestion of the vector generates long overhangs complementary to the USER tails of the amplified gene and promoter fragments. (d) The resulting vector carrying the expression cassette is propagated and purified from *E. coli*. (e) The purified plasmid is linearized by NotI digestion. (f) The linearized construct from the backbone is then transformed into yeast, where the homology regions (UP and DOWN) trigger homologous recombination of the construct into the yeast genome

The primers used for the optimal BenM biosensor design are listed in Table 2.

1. Prepare the PCRs as follows:
 Add to 50 μ L nuclease-free water.
 10 μ L 5 \times Phusion HF Buffer.

2 μL forward primer (10 μM).

2 μL reverse primer (10 μM).

1 μL dNTP mix (10 mM each).

X μL template DNA (typically 1–100 ng of plasmid or synthetic DNA).

1 μL Phusion U Hot Start DNA Polymerase.

2. Run the following PCR program:

98 °C for 30 s.

30 cycles of

98 °C for 10 s.

58 °C for 30 s (or another suitable annealing temperature).

72 °C for 30 s per 1 kb of the PCR product.

72 °C for 5 min.

10 °C hold.

3.2.4 Analysis and Purification of PCR Results

PCR products are analyzed by 1% agarose gel electrophoresis and purified from the gel using the NucleoSpin® Gel and PCR Cleanup from Macherey Nagel, eluting with 50 μL of elution solution. Amplified genes and promoters are stored at -20 °C.

3.3 Vectors Preparation

We use yeast-integrative plasmids from the EasyClone system as vector backbones (*see Note 6*). The biosensor and reporter will ultimately each be cloned into a different integration vector. In this example vector pCfB257 (targeting EasyClone integration site X-3) and pCfB262 (targeting site XII-4) will be prepared.

EasyClone vectors are composed of: pUC plasmid origin of replication and the ampicillin-resistance gene (AmpR), which allow for replication and selection in *E. coli*; NotI USER restriction sites flanking 500 bp homologous regions—named UP and DOWN—for homologous recombination into a target yeast genomic locus; an auxotrophic selectable marker gene (*Kl.LEU2* for pCfB257 and *Sp.HIS5* for pCfB262) to select yeast transformants; and, two yeast terminators ADH1 and CYC1 separated by an SfaAI restriction site. USER-cloned constructs of the biosensor and the reporter are designed to be inserted between the double terminators (Fig. 2c).

3.3.1 Composition of Expression Cassettes

To construct the CCM biosensor in yeast, we originally tested several promoters for controlling the expression of BenM and the yeGFP reporter gene. In the following we illustrate the example for constructing the optimal CCM biosensor using REV1p to control the expression of BenM, and 209bp_CYC1p_BenO_T1 to control the yeGFP reporter gene [16].

3.3.2 Vector Linearization

The backbone linearization is a two-step digestion to create first a double-strand cut and then a single-strand cut to generate long overhangs compatible with USER overhangs (Fig. 2b).

1. Prepare the reaction as follows:
 Add to 50 μL nuclease-free water.
 x μL of EasyClone vector (5 μg).
 5 μL of FastDigest[®] buffer.
 5 μL of FastDigest SfaAI[®] restriction enzyme.
2. Incubate for at least 2 h at 37 °C (*see Note 7*).
3. Purify the plasmid from solution, using NucleoSpin[®] Gel and PCR Clean-up kit from Macherey Nagel, eluting with 50 μL elution buffer.
4. Determine the DNA concentration.
5. Prepare the reactions as follows:
 Eluent (all).
 1 μL of Nb.BsmI/ μg of digested vector.
 x μL of Buffer 3.1 (making up a 1:10 of the total volume).
6. Incubate for 1 h at 65 °C.
7. Purify the digested and nicked vector from the gel, using NucleoSpin[®] Gel and PCR Clean-up from Macherey Nagel. Elute with 50 μL of elution buffer (*see Note 8*).
8. Determine DNA concentration.
9. Store the USER-ready vectors at -20 °C.

3.4 Cloning of Genes and Promoters into Yeast Integration Vectors

The next step is USER cloning of the genes and promoters, generated in Subheading 3.2, into USER-ready integration vectors, generated in Subheading 3.3. Two integration vectors will be created. The first vector will contain BenM under the control of REV1p and will be targeted to EasyClone integration site X-3, whereas the second vector will contain the yeGFP reporter gene driven by 209bp_CYC1p_BenO_T1, and will be targeted to integration site XII-4.

3.4.1 USER Cloning (Fig. 2d)

1. Prepare the USER reaction as follows:
 Add to 6 μL nuclease-free water.
 1 μL of SfaAI/Nb.BsmI-treated vector (pCfB257 or pCfB262 (*see Note 9*)).
 1 μL of promoter fragment for position Promoter 2 (REV1p or reporter promoter).
 1 μL of gene fragment for position Gene 2 (BenM or yeGFP).
 1.2 μL 5 \times Phusion HF buffer.
 0.5 μL USER[™] enzyme.

2. Incubate the mixture in PCR machine at the following conditions:
 - 37 °C for 25 min.
 - 25 °C for 10 min.
 - 20 °C for 10 min.
 - 15 °C for 10 min.
 - 10 °C pause.

3.4.2 Plasmid Amplification

The reaction mix is transformed into DH5 α competent *E. coli* cells (kept at -80 °C)

1. Cool the USER reaction tubes on ice and add 50 μ L of competent cells.
2. After 10 min on ice, perform heat shock at 42 °C for 45 s and place the tubes on ice for 1–2 min.
3. Plate the cells on LB plates with ampicillin and incubate at 37 °C overnight.

3.4.3 Vector Verification and Purification

The correct cloning of gene and promoter fragments into the EasyClone vector is established by PCR on bacteria colonies employing standard primers (*see Note 10*) that amplify the entire expression cassette (Table 3).

The colony PCR is performed as following:

1. Mix the following in a PCR tube:
 - 5 μ L 2xOneTaq Master Mix polymerase.
 - 1 μ L forward verification primer (10 μ M) ID224.
 - 1 μ L reverse verification primer (10 μ M) ID225.
 - 3 μ L nuclease-free water.
2. Add a small amount of *E. coli* colony biomass (it is enough to touch the colony with a pipette tip) to the PCR tube.
3. Run the following PCR program:
 - 94 °C for 3 min.
 - 35 cycles of
 - 94 °C for 20 s.
 - 50 °C for 30 s (or another suitable annealing temperature).
 - 68 °C for 1 min per 1 kb of the PCR product.
 - 68 °C for 5 min.
 - 10 °C pause.
4. Analyze the PCR reactions on 1% agarose gel.

5. The colony corresponding to the correct gel band size (*see Note 11*) is inoculated into 3 mL LB medium with ampicillin and cultivated at 37 °C overnight.
6. The plasmid is purified from the culture using NucleoSpin® kit from Macherey-Nagel.
7. The constructed vector is sequence validated using standard Sanger sequencing and DNA alignment tools.

3.5 Genomic Integration of Expression Constructs

Genomic integration ensures stable expression and reduced cell-to-cell variation in fluorescence output. Therefore, it is advised to integrate both the sensor and reporter constructs into defined loci (*see Note 12*). In this example the sensor construct is integrated into EasyClone site X-3, whereas the reporter construct is integrated into site XII-4.

3.5.1 Expression Vector Linearization

NotI restriction sites flanking UP and DOWN regions in the vector permit the release of the expression cassette needed for homologous recombination in yeast (Fig. 2e).

1. The reaction is set up as follows:
 - x μ L of expression vector (5 μ g).
 - 2 μ L of FastDigest® buffer.
 - X μ L of FastDigest NotI® (use 0.2 μ L per 1 μ g DNA).
 - Add to 20 μ L nuclease-free water.
2. Incubate the reaction at 37 °C for 1 h (*see Note 13*).
3. If the fragment is not purified from agarose, deactivate the enzyme by incubating the reaction at 65 °C for 15 min.
4. The digested vector can be stored at –20 °C for future use.

3.5.2 Transformation of Expression Vector into Yeast (Fig. 2f)

1. Streak CEN.PK113-5A on YPD agar and incubate at 30 °C for 1–2 days.
2. Inoculate 3–5 mL liquid YPD with CEN.PK113-5A and incubate at 30 °C O/N.
3. In the morning: Measure the OD₆₀₀ of the O/N culture. To a culture flask containing 25–50 mL YPD (or media of choice) (*see Note 14*) add cells to give a starting density of approximately 5×10^6 cells/mL. Incubate the flask on a shaker at 30 °C and 200 rpm for two population doublings (typically 4–5 h).
4. When the culture has reached a density of at least 2×10^7 cells/mL harvest the cells by centrifugation, 3000 rpm (1630 $\times g$ in a 16.2 cm radius rotor) for 5 min.
5. Remove the supernatant and resuspend the cells in 25 mL sterile H₂O, centrifuge 3000 rpm (1630 $\times g$) for 5 min and discard the supernatant.

6. Resuspend the cells in 1 mL sterile H₂O and transfer to an Eppendorf tube.
7. Centrifuge 3000 rpm (1630 × *g*) for 30 s and discard the supernatant.
8. Resuspend cells in a total volume of 1 mL H₂O.
9. For each transformation, including a water control, transfer the volume of cell suspension corresponding to 1×10^8 cells (typically 100 μL) to an Eppendorf tube.
10. Centrifuge at 3000 rpm (1630 × *g*) for 30 s and discard supernatant.
11. Make a transformation mix for the planned number of transformations plus one extra and keep it on ice. For each transformation prepare as follows:
 - 240 μL of PEG3350.
 - 36 μL of LiAc 1 M.
 - 10 μL of ssDNA (10 mg/mL).
 - X μL DNA to transform.
 - 74–x μL of MilliQ water.
12. First add X μL of linearized vector corresponding to 300–700 ng (typically 1–10 μL) directly to the cells. Then make a master mix of PEG, LiAc, and ssDNA and add 360–X μL for each transformation. Resuspend the cells by vortex mixing vigorously or pipetting up and down.
13. Incubate the tubes at 42 °C for 40 min.
14. Centrifuge at 3000 rpm (1630 × *g*) for 30 s and remove the transformation mix with a pipette.
15. For selection of amino acid markers: resuspend cells in 100 μL H₂O, and plate on SC-his-leu plates (or markers of choice) and incubate at 30 °C.
16. Check the transformation plate after 2–3 days.
17. Pick a chosen number – we recommend 8 – of colonies and replicate them on selective media. Incubate for 1–2 days at 30 °C.

3.5.3 Genome Integration Verification

The growth of transformants on selective media is not definitive proof of integration in a specific locus. To verify correct genome integration of the linearized fragments, carry out yeast colony PCR as follows:

1. Take a small amount of colony material and resuspend it in 50 μL of 20 mM NaOH.
2. Incubate for 10 min at 100 °C.
3. Spin down the debris and keep the supernatant (DNA).

4. Set up the following genotyping PCR:
 - 6 μL water.
 - 2 μL 2xOneTaq Master Mix polymerase.
 - 1 μL primer 1 (10 μM).
 - 1 μL primer 2 (10 μM).
 - 2 μL of DNA.
5. Run the following PCR program:
 - 94 $^{\circ}\text{C}$ for 1 min.
 - 35 cycles of
 - 94 $^{\circ}\text{C}$ for 20 s.
 - 50 $^{\circ}\text{C}$ for 30 s (or another suitable annealing temperature).
 - 68 $^{\circ}\text{C}$ for 1 min/kb of the PCR product.
 - 68 $^{\circ}\text{C}$ for 7 min.
 - 10 $^{\circ}\text{C}$ pause.
6. Analyze the samples on 1% agarose gel (*see* Table 4).

3.6 Flow Cytometry

In this protocol, biosensor performance is determined using flow cytometry. In short, each strain is pre-cultured individually, and then subcultured both in control medium and in medium with added inducer followed by flow cytometry analysis. Both the level of background fluorescence and the fold induction (fluorescence intensity in the induced state divided by fluorescence intensity in the non-induced state) are important parameters to determine the performance of the biosensor design.

There are two reasons a flow cytometer is used for this purpose. First, the reporter promoter activity of the biosensor design illustrated in this protocol is low, and therefore characterization requires sensitive apparatus to detect the fluorescent signal. Second, flow cytometry allows for single-cell measurements of fluorescent intensity, which allows for insight in the distribution of the signal within the population (Fig. 3).

3.6.1 Induction Media Preparation

It is advisable to prepare pre-mineral medium as described in Sub-heading 2.1 (*see* Note 15). The reason for this is that some inducers, such as CCM, are unstable, and should only be added to the medium on the day the medium is needed.

For the preparation of 100 mL of the induction medium, dissolve 20 mg CCM in 90 mL of pre-mineral medium with a magnetic stirrer, adjust the pH to 4.5 (*see* Note 16) and finally adjust the final volume to 100 mL using deionized water, filter-sterilize and use the same day (*see* Note 17).

For 100 mL of control mineral medium, the recipe is the aforementioned but without the addition of CCM.

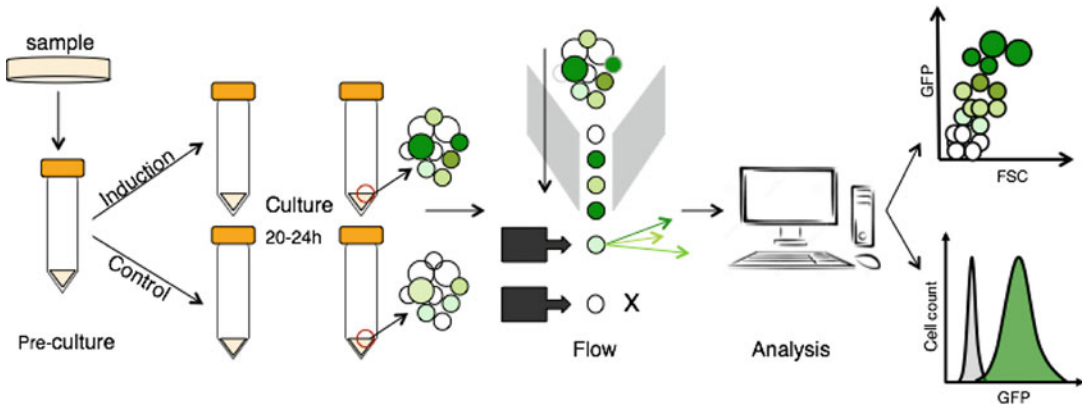


Fig. 3 Flow cytometry analysis of biosensor performance. The yeast strain harboring the sensor and reporter is inoculated for a 24 h pre-culture; subsequently the cells are subcultured into two fresh cultivations either in the absence or presence of the ligand and incubated for another 20–24 h. The cells are diluted in PBS and subjected to flow cytometry for single cell recording of fluorescence level. The output data are exported and more thoroughly analyzed with the use of external software

3.6.2 Culturing Strains and Induction of Biosensor

Ultimately, the performance of each strain is judged by fluorescence measurements of the culture grown in mineral medium with and without inducer. When analyzing a large number of strains, it is advisable to use 96-well plates both for the pre-culture as well as the subculturing step. In addition to the strains containing the biosensor design(s) of interest, it is recommended to analyze the following control strains:

1. WT CEN.PK strain: to assess auto-fluorescence.
2. No sensor, reporter-only control strain: to assess background expression of the reporter gene in the absence of the biosensor.

Day 1: Inoculate 3 single colonies of each strain into 3 different wells of a 96-well plate containing 150–500 μL SC-his-leu per well. Incubate at 250 rpm, 30 $^{\circ}\text{C}$ for 16–20 h.

Day 2: Subculture the strains 1:100 by transferring 5 μL of the overnight culture to both 500 μL control mineral medium and 500 μL induction mineral medium. Incubate at 250 rpm, 30 $^{\circ}\text{C}$ for 20–24 h.

Day 3: After 20–24 h growth, transfer 30 μL of each culture to 150 μL PBS in a 96-well plate right before flow cytometry analysis.

3.6.3 Flow Cytometry and Data Analysis

1. Analyze each culture by flow cytometry with a 488 nm laser for validation of single strains.
2. For each biological replicate of each strain, record 10,000 single-cell events by gating in the forward scatter (FSC) and side scatter (SSC) channels.
3. Export .FSC files.

4. Data can be analyzed using for example FlowJo software (TreeStar Inc.).
5. Calculate the mean fluorescence intensity (MFI) of each biological replicate.
6. Calculate the fold change induction by dividing the MFI of the induced (ON) state by the MFI of the control (OFF) state.
7. Calculate the average and standard deviation of each strain of the MFIs and fold-change induction based on the biological replicates ($n = 3$).

3.6.4 Determining the Operational Range and Specificity of the Biosensor

In order to apply a biosensor for screening cell factory performance, the operational range and specificity of the biosensor needs to be determined. In order to do this, the same steps as in Subheadings 3.6.1–3.6.3 are followed with the following modifications:

1. To determine the operational range: grow each strain in a series of different media that contain increments in the concentration of inducer, for examples 0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4 mM CCM. Plot the average of the mean fluorescence intensity against the inducer concentration to obtain a response curve (Fig. 4a).
2. To determine the specificity: grow each strain in a series of media that have been supplemented with different inducers (*see Note 18*), such as pathway intermediates or chemicals that are structurally similar to the compound of interest. In the case of CCM, malonic acid, protocatechuic acid, fumaric acid, and succinic acid can be tested. Plot the average of the mean fluorescence intensity for both control medium, media with different inducers, as well as medium with CCM (Fig. 4b).

4 Notes

1. A minimal transcription unit (TU) in prokaryotes is composed of several genetic elements: a regulatory region termed a promoter, a transcription start site (TSS), open-reading frames (ORFs) encoding one or more genes, and a transcription termination sequence (TTS) termed a terminator. Importantly, the regulatory region contains the promoter where RNA polymerase and transcription factors (TFs) bind in order to modulate the activity of the promoter. Also, in many cases, prokaryotic transcription initiation requires proteins known as sigma factors (s) that enable proper promoter recognition by RNA polymerase. In general transcription regulation can have a negative effect on promoter activity when the TF, also known as a repressor, binds to the promoter and thereby physically

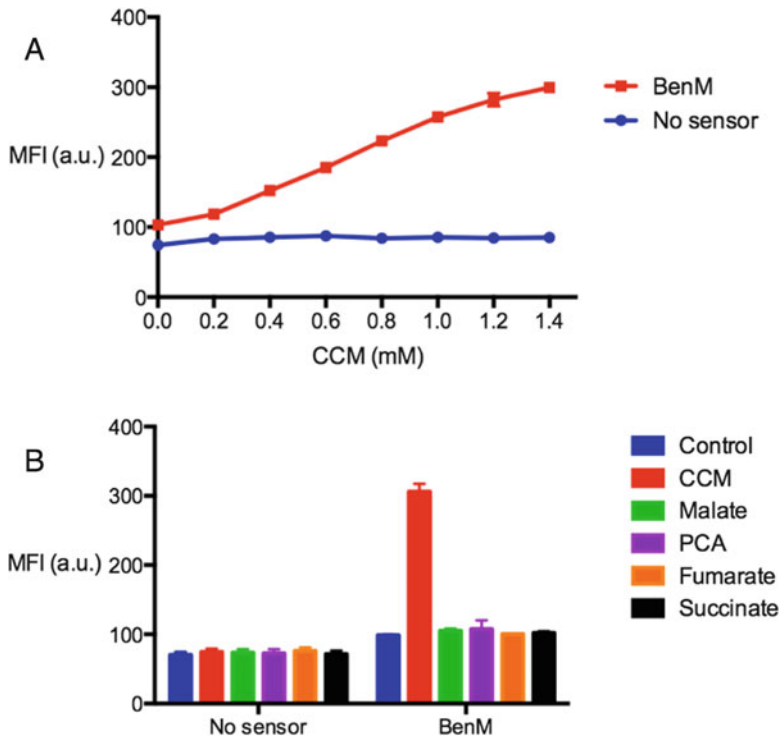


Fig. 4 Operational range and specificity of CCM biosensor. (a) Response function for a yeast harboring the optimal CCM biosensor design (BenM) as well as a control strain only expressing yeGFP from 209bp_CYC1p_-*BenO_T1* (No sensor). The mean fluorescence intensity (MFI) was measured by flow cytometry 24 h after subculturing in the presence of different concentrations of CCM. (b) Specificity of CCM biosensor was determined by measuring yeGFP expression 24 h after subculturing in the presence of various dicarboxylic acids (1.4 mM) and in control medium

interferes with RNA polymerase binding to the promoter (Fig. 1a). In contrast, a positive effect on promoter activity can occur when the regulator, known as activator, bind to the promoter's upstream region to recruit RNA polymerase and initiate transcription (Fig. 1b), or when the repressor undergoes a conformational change disabling its DNA binding and thereby relieving repression (Fig. 1c). In the case of BenM, the TF is constitutively bound to DNA, and only in the presence of CCM the BenM tetramer changes conformation enabling a shift in its binding to the operators. This shift is believed to allow access to DNA binding of the RNA polymerase and thereby CCM-induced activation of expression [18] (Fig. 1d). In native hosts, TFs work jointly and a regulatory region can be occupied by several TFs. Moreover, different sites are able to recruit the same TF, and different TFs can recognize similar sites. Theoretically, the regulatory effect on expression depends on the TF concentration and TF-operator binding affinity: to function, strong sites work with a lower amount, likely for

fine-tuning regulation of important genes, while weak sites require high concentrations of TFs, such as global TFs that are known to be less specific. Furthermore, there are TFs with a dual regulatory role, functioning as activators and repressors at the same time. This is a common theme in sugar catabolism [23].

2. Our design has been tested and validated for other transcriptional activators of the LTTR superfamily as well. Here we just present the protocol for the transplantation of BenM.
3. For the malonic acid biosensor described in our research paper, we first integrated the *SpMAEI* gene, coding for a dicarboxylic acid plasma membrane transporter, to allow for uptake of malonic acid.
4. In case the bacterial sequence has the alternative start codon GTG, change this to ATG.
5. Other fluorescent proteins can be used as well. We also have good experience with sfGFP.
6. Any given integrative plasmid backbone can be used.
7. It is advised to incubate overnight for efficient digestion.
8. The quality of the digest can be tested by gel electrophoresis analysis.
9. Use vector to insert molar ratio 1:3.
10. Standard verification primers are suitable for the amplification of fragments no longer than 5 kb.
11. The sizes are 2025 bp for the construct REV1p-BenM; 1109 bp for 209bp_CYC1p_BenO_T1-yeGFP.
12. Integration of the reporter is in our experience most crucial. We have seen comparable results when the sensor is expressed from a centromeric plasmid.
13. Optional step to confirm linearization on the gel and if desired purify the correct fragment from the gel using NucleoSpin[®] Gel and PCR Clean-up from Macherey Nagel. NotI-digested fragment sizes are 6099 bp for the REV1pr-BenM expression cassette and 4102 bp for the 209bp_CYC1p_BenO_T1-yeGFP. Be aware of the other backbone fragment of 2.8 kb.
14. Adjust depending on the number of transformations needed. 5 mL culture is typically needed per transformation.
15. Add amino acids depending on the auxotrophies of your strain. Since our strains are auxotrophic for tryptophan, we add tryptophan here.
16. Only protonated acids will passively diffuse across the yeast cell membrane. In order to characterize the CCM biosensor we used medium with pH 4.5 in order to leave most of acid in the

protonated state (CCM $pK_a = 3.87$). The pH of the medium should be tuned with the pK_a of your inducer in case uptake of the chemical relies on passive diffusion.

17. It is essential to add and dissolve CCM in mineral medium on the same day the medium is used. We noticed decreased induction of the sensor-reporter when using medium that has been standing for more than a few days. Other inducers might be more stable, though.
18. Use the same molarity of these inducers as the most relevant molarity that was used for the compound of interest (typically the highest molarity).

References

1. Zhang F, Keasling J (2011) Biosensors and their applications in microbial metabolic engineering. *Trends Microbiol* 19(7):323–329. doi:10.1016/j.tim.2011.05.003
2. Zhang J, Jensen MK, Keasling JD (2015) Development of biosensors and their application in metabolic engineering. *Curr Opin Chem Biol* 28:1–8. doi:10.1016/j.cbpa.2015.05.013
3. Rogers JK, Taylor ND, Church GM (2016) Biosensor-based engineering of biosynthetic pathways. *Curr Opin Biotechnol* 42:84–91. doi:10.1016/j.copbio.2016.03.005
4. Zhang C, Wei ZH, Ye BC (2013) Imaging and tracing of intracellular metabolites utilizing genetically encoded fluorescent biosensors. *Biotechnol J* 8(11):1280–1291. doi:10.1002/biot.201300001
5. Liu D, Evans T, Zhang F (2015) Applications and advances of metabolite biosensors for metabolic engineering. *Metab Eng* 31:15–22. doi:10.1016/j.ymben.2015.06.008
6. Park M, Tsai SL, Chen W (2013) Microbial biosensors: engineered microorganisms as the sensing machinery. *Sensors (Basel)* 13(5):5777–5795. doi:10.3390/s130505777
7. Mahr R, Frunzke J (2016) Transcription factor-based biosensors in biotechnology: current state and future prospects. *Appl Microbiol Biotechnol* 100(1):79–90. doi:10.1007/s00253-015-7090-3
8. Stanton BC, Nielsen AAK, Tamsir A, Clancy K, Peterson T, Voigt CA (2014) Genomic mining of prokaryotic repressors for orthogonal logic gates. *Nat Chem Biol* 10(2):99–105. doi:10.1038/nchembio.1411
9. Ramos JL, Martinez-Bueno M, Molina-Henares AJ, Tera W, Brennan R, Tobes R (2005) The TetR family of transcriptional repressors. *Microbiol Mol Biol Rev* 69(2):1–31. doi:10.1128/MMBR.69.2.326
10. Baron U, Bujard H (2000) Tet repressor-based system for regulated gene expression in eukaryotic cells: principles and advances. *Methods Enzymol* 327:401–421. doi:10.1016/S0076-6879(00)27292-3
11. Bertram R, Hillen W (2008) The application of Tet repressor in prokaryotic gene regulation and expression. *Microb Biotechnol* 1(1):2–16. doi:10.1111/j.1751-7915.2007.00001.x
12. Aleksandrov A, Schuldt L, Hinrichs W, Simonson T (2009) Tetracycline-Tet repressor binding specificity: insights from experiments and simulations. *Biophys J* 97(10):2829–2838. doi:10.1016/j.bpj.2009.08.050
13. David F, Nielsen J, Siewers V (2016) Flux control at the Malonyl-CoA node through hierarchical dynamic pathway regulation in *Saccharomyces cerevisiae*. *ACS Synth Biol* 5(3):224–233. doi:10.1021/acssynbio.5b00161
14. Li S, Si T, Wang M, Zhao H (2015) Development of a synthetic Malonyl-CoA sensor in *Saccharomyces cerevisiae* for intracellular metabolite monitoring and genetic screening. *ACS Synth Biol* 4(12):1308–1315. doi:10.1021/acssynbio.5b00069
15. Teo WS, Chang MW (2015) Bacterial XylRs and synthetic promoters function as genetically encoded xylose biosensors in *Saccharomyces cerevisiae*. *Biotechnol J* 10(2):315–322. doi:10.1002/biot.201400159
16. Skjoedt ML, Snoek T et al (2016) Engineering prokaryotic transcriptional activators as metabolite biosensors in yeast. *Nat Chem Biol* 12(11):951–958
17. Curran KA, Leavitt JM, Karim AS, Alper HS (2013) Metabolic engineering of muconic acid

- production in *Saccharomyces cerevisiae*. *Metab Eng* 15(1):55–66. doi:[10.1016/j.ymben.2012.10.003](https://doi.org/10.1016/j.ymben.2012.10.003)
18. Collier LS, Gaines GL, Neidle EL (1998) Regulation of benzoate degradation in *Acinetobacter* sp. strain ADP1 by BenM, a LysR-type transcriptional activator. *J Bacteriol* 180(9):2493–2501
 19. Bundy BM, Collier LS, Hoover TR, Neidle EL (2002) Synergistic transcriptional activation by one regulatory protein in response to two metabolites. *Proc Natl Acad Sci U S A* 99(11):7693–7698. doi:[10.1073/pnas.102605799](https://doi.org/10.1073/pnas.102605799)
 20. Jensen NB, Strucko T, Kildegaard KR, David F, Maury J, Mortensen UH, Forster J, Nielsen J, Borodina I (2014) EasyClone: method for iterative chromosomal integration of multiple genes in *Saccharomyces cerevisiae*. *FEMS Yeast Res* 14(2):238–248. doi:[10.1111/1567-1364.12118](https://doi.org/10.1111/1567-1364.12118)
 21. McIsaac RS, Gibney PA, Chandran SS, Benjamin KR, Botstein D (2014) Synthetic biology tools for programming gene expression without nutritional perturbations in *Saccharomyces cerevisiae*. *Nucleic Acids Res* 42(6):1–8. doi:[10.1093/nar/gkt1402](https://doi.org/10.1093/nar/gkt1402)
 22. Mikkelsen MD et al (2012) Microbial production of indolylglucosinolate through engineering of a multi-gene pathway in a versatile yeast expression platform. *Metab Eng* 14(2):104–111
 23. Balleza E, López-Bojorquez LN, Martínez-Antonio A, Resendis-Antonio O, Lozada-Chávez I, Balderas-Martínez YI, Encarnación S, Collado-Vides J (2009) Regulation by transcription factors in bacteria: beyond description. *FEMS Microbiol Rev* 33(1):133–151. doi:[10.1111/j.1574-6976.2008.00145.x](https://doi.org/10.1111/j.1574-6976.2008.00145.x)

A Capture-SELEX Strategy for Multiplexed Selection of RNA Aptamers Against Small Molecules

Lasse H. Lauridsen, Holger B. Doessing, Katherine S. Long,
and Alex T. Nielsen

Abstract

In vitro selection of aptamers that recognize small organic molecules has proven difficult, in part due to the challenge of immobilizing small molecules on solid supports for SELEX (Systematic Evolution of Ligands by Exponential Enrichment). This study describes the implementation of RNA Capture-SELEX, a selection strategy that uses an RNA library to yield ligand-responsive RNA aptamers targeting small organic molecules in solution. To demonstrate the power of this method we selected several aptamers with specificity towards either the natural sweetener rebaudioside A or the food-coloring agent carminic acid. In addition, Bio-layer interferometry is used to screen clonal libraries of aptamer candidates and is used to interrogate aptamer affinity. The RNA-based Capture-SELEX strategy described here simplifies selection of RNA aptamers against small molecules by avoiding ligand immobilization, while also allowing selection against multiple candidate targets in a single experiment. This makes RNA Capture-SELEX particularly attractive for accelerated development of RNA aptamers targeting small metabolites for incorporation into synthetic riboswitches and for analytical biosensors.

Key words SELEX, RNA aptamer, Small molecules, Next-generation sequencing, Bio-layer interferometry

1 Introduction

Aptamers are single-stranded oligonucleotides that adopt defined tertiary structures, allowing them to recognize and bind specific targets with high affinity and selectivity. These properties make aptamers particularly useful as biosensor elements and components of riboswitches that regulate gene expression. Most aptamers are discovered through an in vitro selection process called SELEX [1, 2]. A combinatorial library, typically containing 10^9 – 10^{15} unique sequences, is incubated with the target of interest, and unbound sequences are discarded to enrich the pool for putative aptamers. The enriched pool is then amplified and may be subjected to repeated rounds of selection until the pool reaches the desired

affinity. This process may eventually yield individual aptamers with affinities in the nano- or picomolar range. The fact that the selection is carried out *in vitro* confers a significant advantage over custom antibody production by allowing the use of compounds that are otherwise difficult to target for reasons such as toxicity or rapid biodegradation.

Additionally, RNA aptamers may be further engineered into synthetic riboswitches that can have regulatory functions in the cell [3–5]. The advent of synthetic biology has led to an increased interest in creating strains for the production of value-added compounds, with the term “cell factories” being used to describe strains that have been engineered to produce specific compounds in yields that make them viable on an industrial scale. Metabolic engineering of such strains requires careful optimization of each step in the biosynthetic pathway as well as its interaction with the native metabolism of the host organism. This generally necessitates the use of directed evolution, targeted mutagenesis, and the construction of recombinant libraries, followed by high-throughput screening for the desired phenotype. The biosynthesis of some compounds may be measured from their color or auto-fluorescence, but most are not readily detectable and require time-consuming extraction and analytical methods, which significantly lowers throughput. High-throughput screens of such compounds are therefore often based on enzyme-coupled assays, growth complementation, or biosensor-driven reporter production [6].

Although some biological compounds have naturally occurring enzymes or regulators, such as transcription factors that may be engineered into biosensors, many compounds have no known regulator, prompting researchers to look for custom-tailored solutions. Raising aptamers specific to the target compound(s) and subsequently engineering them into riboswitches in principle enables the regulation of gene expression in response to the target compounds. Gene regulation in response to higher titers of industrially interesting compounds has been used recently to select for tryptophan production phenotypes in selection assays [7].

Creating novel riboswitches is challenging and the most straightforward approach is to build upon the structure-switching properties of an existing RNA aptamer. Aptamers typically undergo a conformational change upon binding to their targets [8, 9]. When an aptamer is part of a larger transcript this rearrangement may either break or establish new interactions with its flanking sequences and thus mask or unmask nearby elements, such as ribosomal binding sites, (anti-) terminators, or splicing sites. The overall effect is an impact on gene expression that is either increased or decreased in response to the concentration of the target molecule. Rational design of these interactions is currently not feasible and the chosen aptamer is therefore often recombined with one or more random sequence elements and subsequently screened for the

desired activity in vivo [10–12]. By using an aptamer with a pre-established conformational rearrangement that is triggered by target binding, the prospect of realizing a synthetic riboswitch with the desired function is improved.

Capture-SELEX is a variant of the in vitro selection process, which is well suited for isolating aptamers against small soluble target molecules [13]. In classical SELEX, aptamer candidates are selected by passing the oligonucleotide library over the immobilized target molecule. However, in Capture-SELEX the oligonucleotide library itself is immobilized by annealing to a bead-bound capture probe and aptamer candidates are eluted with a solution of the solvated target. Any library members that dissociate from the probe due to gross conformational changes upon binding the target ligand are then collected, amplified, and used in the next round of selection. This process is repeated until the pool has been sufficiently enriched for putative aptamers. As the target remains solubilized throughout the selection process, the elution of aptamer candidates lends itself well to multiplexing using several targets. This method was first reported to yield DNA aptamers against kanamycin A [13]. Here, we present a variant of Capture-SELEX, which uses an RNA library to yield RNA aptamers against small organic molecules. Following eight rounds of RNA Capture-SELEX, we have successfully selected aptamers against rebaudioside A, a natural sweetener, as well as carminic acid, a food coloring agent [14] (Fig. 1).

2 Materials and Methods

Preferably, perform all the following experiments in an area designated for RNA work, remember to wear gloves, and to all possible extent use RNase-free reagents. We recommend using filter tips to prevent RNase contamination and carry-over of library between samples.

2.1 Library Preparation

1. 10× hybridization buffer (200 mM Tris-HCl, pH 8.0, 500 mM NaCl, 10 mM EDTA).
2. CASE001 DNA template library with central capture probe region (desalted 5'-AGA-TTG-CAC-TTA-CTA-TCT(N)₄₀GAT-CGAGCC-TCA(N)₁₀AAT-TGA-ATA-AGC-TGG-TAT-CCT-ATA-GTG-AGT-CGT-ATTAG-3', Integrated DNA Technologies).
3. CASE003 reverse primer (desalted, 5'-CTA-ATA-CGA-CTC-ACT-ATA-GGA-TAC-CAG-CTT-ATT-CAATT-3).
4. TranscriptAid T7 High Yield Transcription kit (1 mL, Thermo Scientific).
5. 0.22 μm Corning Costar Spin-X filter (Sigma-Aldrich).

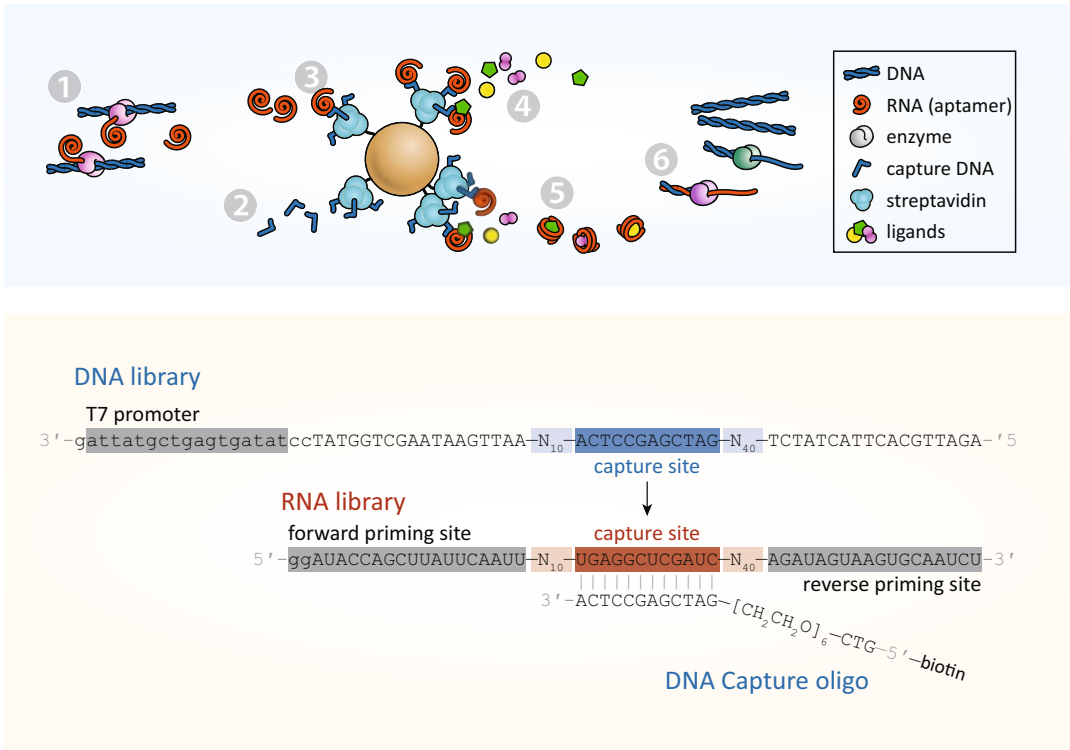


Fig. 1 Rationale behind RNA Capture-SELEX. (*Upper panel*): An overview of the RNA Capture-SELEX approach. (1) An RNA library is generated from a DNA template by in vitro transcription. (2) A biotinylated capture oligonucleotide is immobilized onto paramagnetic streptavidin beads. (3) The RNA library is docked onto the immobilized capture oligonucleotide. (4) Following extensive washing of the beads, one or more ligands are added. (5) Putative structure-switching aptamers, i.e., RNAs that bind the ligand and are released from the beads in the process, are collected from the supernatant. (6) If further rounds of selection are required, the eluted RNA is used as a template for reverse transcription and PCR to regenerate the DNA library. (*Lower panel*): Design of the libraries used in this protocol. The DNA library is transcribed to yield the corresponding RNA library. The biotinylated capture oligomer binds the RNA library via its capture site flanked by 10 upstream and 40 downstream randomized positions (N10, N40). The forward and reverse priming sites are used for reverse transcription and PCR [14]

6. TRIzol (Sigma-Aldrich).
7. Chloroform.
8. Isopropanol.
9. RNase-free water.
10. DNase I (1.25 units/μg DNA).
11. Bioanalyzer RNA Nano 6000 kit (Agilent).
12. Qubit RNA HS kit (Life Technologies).
13. RiboLock RNase inhibitor (40 U/μL, Thermo Fisher Scientific).

2.2 Capture-SELEX

1. M-270 Streptavidin Dynabeads.
2. Magnetic rack for 1.5 mL Eppendorf tubes.

3. 0.2 M NaOH in RNase-free water.
4. 0.1 M NaCl in RNase-free water.
5. Binding and washing buffer, B&W (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl, and 0.005% (v/v) Tween-20).
6. CASE004 capture oligonucleotide (HPLC-purified 5'-biotin-GTC-[hexaethylene glycol]-GATCGAGCCTCA-3', IDT).
7. CS buffer (20 mM Tris-HCl pH 7.4, 2 mM MgCl₂, 5 mM KCl, 1 mM CaCl₂, 100 mM NaCl and 0.005% (v/v) Tween-20).
8. Target ligands (0.33 mM carminic acid, 0.5 mM rebaudioside A, 0.5 mM l-5-hydroxytryptophan).

2.3 Quantitation of Eluates by RT-qPCR

1. RevertAid reverse transcriptase (200 U/μL).
2. EvaGreen qPCR dye, 20× in water (Biotium).
3. ROX reference dye, 1 μM in RNase-free water.
4. Phusion HF II PCR enzyme (2 U/μL).
5. Phusion HF buffer, 5×.
6. dNTPs, 10 mM.
7. CASE003 reverse primer (desalted, 5'-CTA-ATA-CGA-CTC-ACT-ATA-GGA-TAC-CAG-CTT-ATT-CAATT).
8. CASE002 forward primer (desalted, 5'-AGA-TTG-CAC-TTA-CTA-TCT).
9. Bioanalyzer High Sensitivity DNA kit (Agilent).
10. Amicon Ultra-0.5 30 K spin filters.

2.4 Blunt-End Cloning of Aptamer Pools

1. CloneJet PCR cloning kit.
2. XL-Blue Subcloning-Grade Competent Cells.
3. Agencourt AMPure XP.
4. Qiagen Plasmid Plus 96 kit.

2.5 Monitoring RNA Release by Biolayer Interferometry

1. Black 96-well microplates (Greiner Bio-One).
2. Dip-and read Streptavidin (SA) Biosensors.
3. Regeneration buffer: 50 mM NaOH, 1 M NaCl.
4. Agencourt AMPure XP.
5. Direct-zol RNA miniprep kit.

3 Methods

Prepare all solutions using RNase-free water and analytical-grade reagents. Use filtered tips and gloves to prevent RNase contamination.

**3.1 Library Prep for
First Round of Capture
SELEX**

1. Using filter pipet tips, add 20 μL 10 \times hybridization buffer, 8 μL 1 mM (8 nmol) CASE003, 40 μL 100 μM CASE001 (4 nmol) and 132 μL RNase-free water into a 1.5 mL RNase-free tube and mix by vortexing.
2. Incubate at 80 $^{\circ}\text{C}$ for 5 min and let cool at room temperature for 30 min.
3. Place at 5 $^{\circ}\text{C}$ for 10 min prior to setting up the T7 transcription reaction.
4. To the hybridization reaction add 200 μL 5 \times TranscriptAid T7 buffer, 100 μL 100 mM rATP, 100 μL 100 mM rCTP, 100 μL 100 mM rUTP, 100 μL 100 mM rGTP, 100 μL T7 enzyme mix, and 100 μL RNase-free water.
5. Incubate at 30 min at 37 $^{\circ}\text{C}$ on a heat block (*see Note 1*).
6. Set centrifuge to cool to 4 $^{\circ}\text{C}$.
7. Remove the DNA template by adding 200 μL DNase I to the transcription reaction and incubate for 15 min at 37 $^{\circ}\text{C}$.
8. Split the reaction into two 650 μL aliquots and remove the precipitate by passing both through a Costar Spin-X 0.22 μm spin filter (4 min at 1000 $\times g$) (*see Note 2*).
9. Divide the filtrates into 5 RNA free tubes (260 μL each).
10. To each vial, add 1/10 volume of 3 M NaOAc, pH 5.2 (supplied with the T7 kit).
11. Add three volumes (780 μL) of Trizol and mix by pipetting.
12. Add one volume (260 μL) of chloroform and shake 15 s by hand.
13. Spin at 12,000 $\times g$ for 2 min at 4 $^{\circ}\text{C}$ to separate phases.
14. Transfer the upper aqueous phase into a new RNase-free tube.
15. Add one volume (260 μL) of chloroform and repeat shaking and phase separation.
16. Precipitate the RNA by adding one volume of 100% isopropanol to the aqueous phase.
17. Incubate for 10 min at RT.
18. Spin down precipitate at 14,000 $\times g$ for 20 min at 4 $^{\circ}\text{C}$. (The precipitated RNA may be stored like this at -20°C .)
19. Remove the supernatant.
20. Carefully wash the RNA pellet with 600 μL ice-cold 70% ethanol.
21. Spin at 14,000 $\times g$ for 5 min and remove the supernatant.
22. Repeat **steps 21–22**.
23. Air-dry the RNA pellets for 10 min (*see Note 3*).

24. Redissolve pellets in $5 \times 50 \mu\text{L}$ RNase-free water. Combine and store at -80°C .
25. The RNA pool is analyzed using the Agilent Bioanalyzer RNA 6000 nano kit.
26. Prepare an RNA chip as described by the manufacturer.
27. Make $20 \mu\text{L}$ 10-, 100- and 1000-fold dilutions of the RNA pool and run them on the Agilent bioanalyzer to ensure correct peak size and absence of aborted RNA transcripts (*see Note 4*).
28. The RNA quantitation is verified by Qubit using the RNA HS kit.

3.2 Bead Preparation for Capture SELEX Round 1

3.2.1 Coat Beads with Docking Oligonucleotide

1. Take out 10^9 M-270 Streptavidin Dynabeads ($1538 \mu\text{L}$).
2. Place the beads on a magnetic separator and remove the supernatant.
3. Wash the beads in 1 mL 0.1 NaOH, 0.05 NaCl, 0.005% (v/v) Tween-20.
4. Place on magnetic separator and remove supernatant.
5. Repeat washing **steps 3–4**.
6. Repeat washing **steps 3–4** with 0.1 NaCl, 0.005% (v/v) Tween-20.
7. Wash beads three times with $500 \mu\text{L}$ binding and washing buffer.
8. Resuspend the beads in 0.5 mL $1\times$ B&W buffer.
9. Add 0.5 mL $12 \mu\text{M}$ CASE004 docking oligonucleotide in $1\times$ B&W buffer.
10. Incubate for 1 h at RT with end-over-end nutation.

3.2.2 Remove Unbound Docking Oligonucleotide

1. Wash beads three times with $500 \mu\text{L}$ $1\times$ B&W buffer.
2. Wash beads three times with $500 \mu\text{L}$ with $1\times$ CS buffer.
3. Resuspend beads in 1 mL $1\times$ CS buffer.
4. Coated beads may be stored at 5°C , while preparing the RNA.

3.2.3 Load the RNA

1. Dilute 10 nmol RNA library with RNase-free water and $2\times$ CS buffer to a final volume of $300 \mu\text{L}$ $1\times$ CS buffer.
2. Fold the RNA by denaturing at 80°C for 8 min and cool for 10 min at 5°C .
3. Add $7.5 \mu\text{L}$ RiboLock to the sample.
4. The folded RNA library is added to the CASE004-coated beads and left on a thermomixer at 21°C , 1400 RPM overnight.

3.3 RNA Capture SELEX

3.3.1 Removing Unbound RNA Library

1. Place the RNA/bead mixture on a magnetic separator and remove supernatant. Save the supernatant for later analysis by RT-qPCR.
2. Wash the beads nine times with 500 μL 1 \times CS buffer with Tween.
3. Resuspend beads in 300 μL 1 \times CS buffer.

3.3.2 Elution of Weakly Bound RNA Library (Temperature Step)

1. Incubate the RNA/bead mixture at 28 $^{\circ}\text{C}$ for 15 min in a thermomixer at 1200 RPM.
2. Place heated mixture on magnetic separator.
3. Remove and save supernatant
4. Wash the beads seven times with 500 μL 1 \times CS buffer with Tween.
5. Resuspend beads in 300 μL 1 \times CS buffer

3.3.3 Background Elution of RNA Library (Background Elution Step)

1. Incubate the RNA/bead mixture at 21 $^{\circ}\text{C}$ for 45 min in a thermomixer at 1200 RPM.
2. Place heated mixture on magnetic separator.
3. Wash the beads seven times with 500 μL 1 \times CS buffer with Tween. Resuspend beads in 500 μL 1 \times CS buffer with Tween.

3.3.4 Elution of RNA Library with Target Compounds (Ligand Step)

1. Prepare a 0.5 mM stock solution of the ligand compounds in CS buffer.
2. Place bead mixture on magnetic separator and discard supernatant.
3. Add 300 μL ligand mixture to beads.
4. Immediately incubate ligand/bead mixture at 21 $^{\circ}\text{C}$ for 45 min in a thermomixer at 1200 RPM.
5. Place heated mixture on magnetic separator.
6. Remove and save supernatant for later analysis.

3.3.5 Heat Elution of Remaining RNA Library (Heat Elution Step)

1. Resuspend beads in 300 μL 1 \times CS buffer.
2. Incubate the RNA/bead mixture at 80 $^{\circ}\text{C}$ for 5 min in a thermomixer at 1200 RPM.
3. Place heated mixture on magnetic separator.
4. Remove and save supernatant for later analysis
5. Repeat **step 1–4** two additional times, remembering to save supernatants.
6. Proceed to analyze saved supernatant from temperature-, background-, ligand-, and heat-elution steps by qPCR.

3.4 Quantitation of Eluates by RT-qPCR

3.4.1 Reverse Transcription

Reverse transcription-quantitative PCR (RT-qPCR) of the eluates can be used to monitor the progress of the Capture-SELEX, as well as checking for contaminating DNA.

Prepare reverse transcription reactions using RevertAid Premium Reverse Transcriptase in accord with the manufacturer's instructions, observing the following:

1. Use 1 μL aliquot of each RNA sample for RT-qPCR. A total of seven samples from the Capture SELEX step are set up for reverse transcription:

| | |
|--------------------------------|---------------------------------------|
| 1. Unbound RNA | Quantifies the amount of RNA loaded |
| 2. Unbound RNA (1:999 diluted) | Quantifies the amount of RNA loaded |
| 3. Background elution | Quantifies baseline RNA release |
| 4. Elution | Quantifies ligand-induced RNA elution |
| 5. Heat elution (1st of 3) | Quantifies total RNA on beads |
| 6. Heat elution (2nd of 3) | Quantifies total RNA on beads |
| 7. Heat elution (3rd of 3) | Quantifies total RNA on beads |

2. In addition to the unknown samples make a dilution series of the unselected RNA pool prepared in Subheading 3.1. We suggest preparing stock solutions containing 33 nM—33 f. RNA.
3. Also include a positive control with 100 pM CASE001 DNA, as well as a no template control (NTC) in the RT-qPCR.
4. Set up 14.5 μL pre-hybridization reactions containing 1 μL of unknown, standard or control aliquots, 20 pmol CASE002, and 10 μmol of each deoxynucleotide.
5. Heat the mixture to 65 °C for 5 min and chill on ice.
6. Following annealing of the primer to the RNA, prepare “no reverse transcription” control reactions by dividing each sample into two and substitute reverse transcriptase with water.
7. For the positive reaction mix 7.25 μL hybridization mix, 4 μL 5 \times RT buffer, 20 U RiboLock and 200 U RevertAid Premium reverse transcriptase for a total volume of 20 μL . The control reactions without reverse transcriptase are assembled in the same manner, substituting reverse transcriptase with water.
8. Incubate the reverse transcription reactions at 50 °C for 30 min and inactivate the reactions at 85 °C for 5 min.

3.4.2 Quantitative PCR

Due to the relatively low T_m of the primers, we recommend preparing a custom quantitative PCR (qPCR) mix using Phusion DNA polymerase.

1. Prepare 20 μL PCR reactions using Phusion DNA polymerase in HF buffer in accord with the manufacturer's instructions containing 1 μL reverse transcription reaction (Subheading 3.4.1, step 1 a–h), 10 pmol of primers CASE002 and CASE003, 1 \times EvaGreen and 1 \times ROX reference dye.
2. On the thermocycler, choose the appropriate settings for well volume normalization using ROX and quantitation of PCR amplification using EvaGreen.
3. Run the PCR using this program: 1 min at 98 $^\circ\text{C}$; 40 cycles of: 30 s at 98 $^\circ\text{C}$, 30 s at 51 $^\circ\text{C}$, 30 s at 72 $^\circ\text{C}$ *; 5 min at 72 $^\circ\text{C}$. (* Quantitate the reactions at the end of each extension step.).
4. Verify the formation of the appropriate PCR product by resolving 5 μL of each sample on a 2% 1 \times TAE agarose gel.
5. Confirm that there is appropriate amplification of the positive control reaction and that there is no or negligible amplification in the negative control reactions (NTC with buffer only, and NTCs without reverse transcription).
6. Plot Ct-values obtained from qPCR of the standard curves and use the regression curve to estimate RNA concentration in the Capture SELEX samples.

3.5 Subsequent Rounds of RNA Capture SELEX

The subsequent rounds of RNA Capture SELEX begin with preparation of the RNA library. In our hands, RNA and DNA SELEX libraries tend to create erroneous PCR products if subjected to too many PCR cycles. To generate high-fidelity RNA libraries, the optimal number of PCR cycles therefore has to be empirically determined by small-scale pilot PCR.

3.5.1 Preparing dsDNA from RNA Eluate

1. The entire RNA eluate from Subheading 3.3.4 is pelleted by adding 1/10 vol (30 μL) of 3 M NaOAc, pH 5.2 to the eluate.
2. Add 2.5 μL glycogen @ 20 $\mu\text{g}/\mu\text{L}$ (50 μg total).
3. Add 2.5 vol. (825 μL) EtOH. Mix by pipetting and incubate for 1 h at -80°C .
4. Pellet the RNA by centrifugation for 30 min at 14,000 $\times g$, 4 $^\circ\text{C}$.
5. Wash the RNA pellet twice with 500 μL ice-cold 70% EtOH.
6. Air-dry pellets (2 min).
7. Resuspend in 30 μL DEPC-treated water.
8. Add 10 μL 10 μM CASE002, 5 μL 10 mM dNTPs and 27.5 μL RNase free water to the resuspended RNA and vortex to mix.
9. Heat the mixture to 65 $^\circ\text{C}$ for 5 min and chill on ice.

10. Add 20 μL $5\times$ RT buffer, 2.5 μL RiboLock and 5 μL RevertAid reverse transcriptase to the annealed RNA.
11. Incubate at 50 °C for 30 min and inactivate the reverse transcriptase reactions at 85 °C for 5 min.
12. Store at 5 °C.

3.5.2 Small Scale PCR

1. Use 1.8 μL of the generated cDNA to assemble a 30 μL PCR reaction using the Phusion HF standard protocol.
2. Amplify the cDNA using the following protocol: 1 min at 98 °C, 16 \times (30 s at 98 °C, 30 s at 51 °C, 30 s at 72 °C), 5 min at 72 °C.
3. Take 3 μL aliquots during the last 10 s of the extension phase in cycles 4, 6, 8, 10 and 12 (*see Note 5*).
4. Run aliquots on a Bioanalyzer using the High Sensitivity DNA kit.
5. Choose the number of cycles that correspond to a pure PCR product of correct length for the large-scale PCR.

3.5.3 Large Scale PCR

1. Using the remainder of the cDNA 1500 μL PCR is set up as described in Subheading 3.5.2 and aliquoted into smaller volumes in an unskirted PCR plate.
6. Using the determined optimal number of cycles run the PCR program as mentioned in Subheading 3.5.2.
2. After PCR amplification concentrate dsDNA libraries with Amicon Ultra 30 K centrifugal filters approximately to 100 μL (5 min at 14,000 $\times g$).
3. Quantitate the retentate by Qubit.
4. Aliquots of the DNA should be stored for later analysis by next-generation sequencing.
5. The remainder is used for transcribing the RNA library for the following selection round in accordance with Subheading 3.1 (*see Note 6*) (Fig. 2).

3.6 Blunt-End Cloning of Aptamer Pools

1. To isolate clones for screening, the aptamer pools are blunt-end cloned into the pJET 1.2 vector (Thermo Fisher Scientific) and transformed into XL1-Blue Subcloning-Grade Competent Cells (Agilent) using standard protocols.
2. Pick 96 random colonies and perform colony PCR with pJET1.2 Forward Sequencing Primer (5'-CGACTCACTA-TAGGGAGAGCGGC-3') or pJET1.2 Reverse Sequencing Primer (5'-AAGAACATCGATTTTCCATGGCAG-3').
3. Verify the insertion of the aptamer candidate sequences by gel electrophoresis; correctly inserted aptamer candidates yield a 231 bp PCR product.

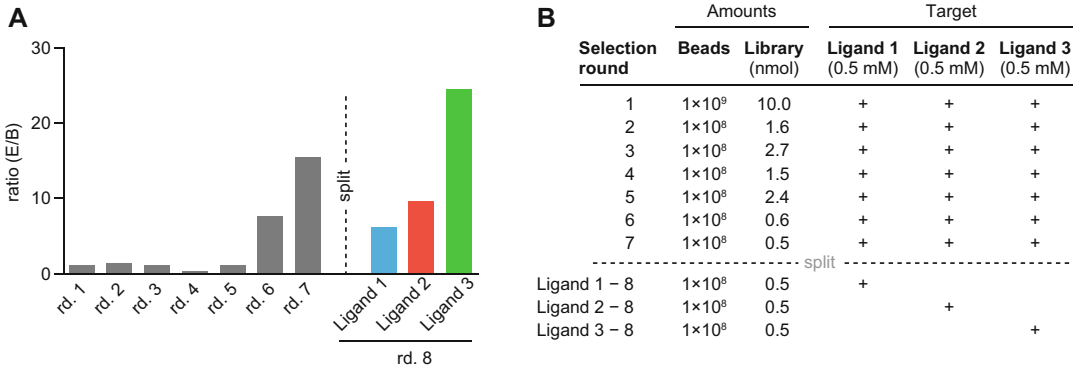


Fig. 2 An example of the progress of capture SELEX as measured by RT-qPCR. **(a)** Eluted RNA over the course of eight rounds of RNA Capture SELEX. Seven rounds of RNA Capture SELEX are performed using a mixture of three ligands to enrich for structure switching RNA. After seven rounds, the mixture can be split and subsequent rounds then feature only a single ligand. The amount of eluted RNA can emphasize which ligand affinity is most enriched for. **(b)** Table depicting loaded amounts RNA library through a standard RNA capture SELEX campaign [14]

4. Make overnight cultures of picked colonies in 1 mL Luria Bertoni broth supplemented with 100 µg/mL ampicillin.
5. Purify plasmids from cultures using the Qiagen Plasmid Plus 96 kit.
6. Store plasmids at -20°C .

3.7 Bio-Layer Interferometry

To identify which of the randomly picked RNA sequences exhibit ligand induced release from the capture oligonucleotide, a screening assay is performed using biolayer interferometry (BLI). The hits obtained from these screens can then be verified in a subsequent assay to determine a release constant (*see Note 7*).

3.7.1 Prepare RNA for BLI Screening

1. Prepare 20 µL PCR reactions using Phusion DNA polymerase in HF buffer in accord with the manufacturer's instructions containing 1 µL plasmid template and 10 pmol of primers CASE002 and CASE003.
2. Verify PCR fragment size by gel electrophoresis.
3. Purify PCR products with AMPure beads according to the manufacturer's guidelines.
4. Check DNA concentration by Qubit or Nanodrop.
5. Use 50 ng (or 1.5 µL) of each PCR product to set up 5 µL T7 transcription reactions using TranscriptAid T7 High Yield Transcription kit in a 96-well unskirted PCR Plate.
6. Run the transcription reactions overnight at 37°C .
7. The completed RNA transcription reactions may be stored at -20°C until needed.

3.7.2 Screening Selected RNA

The bio-layer interferometry (BLI) experiments are performed using the Octet RED96, but can in theory be performed on any equivalent system from ForteBio. Please observe that all BLI experiments are carried out at constant (room) temperature.

1. Dilute the transcriptions reactions in $1\times$ CS buffer to a final volume of 200 μ L immediately before running the bio-layer interferometry assay.
2. Prepare the Dip-and read Streptavidin Biosensors by dipping them in $1\times$ CS buffer for 10 min at 1000 RPM.
3. Activate the biosensors by dipping them in 1 μ M CASE004 in $1\times$ CS buffer for 200 s.
4. Remove unbound CASE004 and record baselines by dipping the sensor tips in $1\times$ CS buffer for 30 s.
5. Load RNA onto the tips by dipping them in the diluted RNA samples for 300 s.
6. Transfer the tips to $1\times$ CS buffer with 0.5 mM of appropriate target and record the release of RNA for 500 s.
7. Regenerate the biosensor tips by alternately dipping three times in regeneration buffer for 5 s and $1\times$ CS buffer for 5 s.
8. Repeat the loading of RNA step as before.
9. Transfer the tips to $1\times$ CS buffer without target ligands and record the release of RNA for 500 s.
10. Subtract background release at 500 s measured in **step 9** from the release measured at 500 s in **step 6** to get an estimate of ligand-induced release.
11. RNA candidates that show significantly increased release as a function of ligand addition are selected for further characterization of release kinetics and selectivity (Fig. 3).

3.7.3 Characterization of RNA Aptamers by BLI

Hits from the initial screening can be characterized by bio-layer interferometry by including an additional dilution series of target. Furthermore, target specificity can be assessed against targets included in the multiplexed selection or other related compounds.

1. Prepare RNA of the selected candidates in accord with Sub-heading 3.7.1, albeit in $10\times$ scale to provide enough RNA for characterization.
2. Purify in-vitro transcription reactions using the Direct-zol RNA miniprep kit according to the manufacturer's description.
3. Analyze the RNA using the Agilent Bioanalyzer RNA 6000 nano kit.
4. Make 20 μ L 10-, 100- and 1000-fold dilutions of each RNA sample and analyze them using the Agilent Bioanalyzer RNA

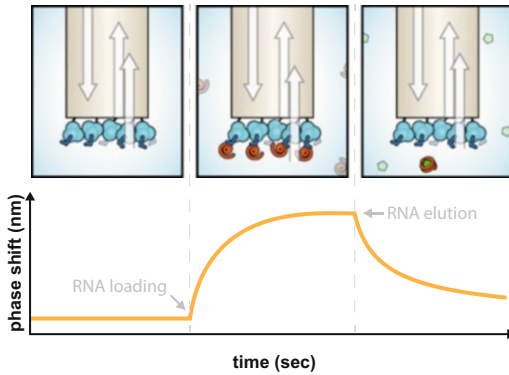
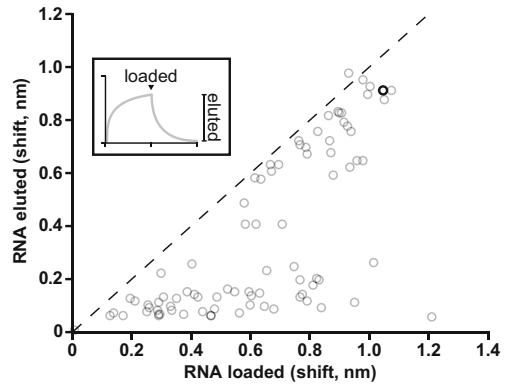
A Characterizing aptamers by BLI**B Loading vs. elution of a set of aptamers**

Fig. 3 Bio-layer interferometry assay examples. **(a)** Response of RNA pools to ligands. **(a, left panel)** Bio-layer interferometry (BLI) is a label-free technique for studying molecular interactions in real-time. An optical fiber coated with a bio-layer is immersed in the analyte solution and white light is directed down the fiber. Light reflected back from the tip of the fiber interferes with light reflected from the interface between the bio-layer (streptavidin and capture oligomer, *blue*) and the surrounding solution, resulting in a distinct interference pattern. **(a, center panel)** When RNA (*red*) binds to the bio-layer the optical thickness at the tip is increased, and the interferometric profile shifts in proportion to the extent of binding. **(a, right panel)** When ligand (*green*) is added it binds to and releases the RNA, thereby reducing the optical thickness. **(a, bottom)** The measurements are presented as a sensorgram indicating the interferometric phase shift as a function of time. **(b)** Screening reveals populations of responsive RNA clones. The amounts of in vitro-transcribed RNA from randomly picked clones from the aptamer library that could be loaded onto BLI sensors in 500 s compared to the extent of elution with 1 mM ligand in 500 s. Dashed lines illustrate the ideal case of complete elution of the loaded RNA. Highlighted data point (*black*) correspond to an example of a putative aptamer [14]. Insert: Illustration of the collection of a set of endpoint values. The arrow marks the addition of ligand

6000 nano kit to ensure correct peak size and absence of aborted RNA transcripts.

5. Determine RNA concentration using Qubit (*see Note 8*).
6. Immediately before setting up the BLI experiment, make 1 μM solutions of each RNA to be tested in $1\times$ CS buffer (*see Note 9*).
7. Prepare the Dip-and-read Streptavidin Biosensors by dipping them in $1\times$ CS buffer for 10 min at 1000 RPM.
8. Activate the biosensors by dipping them in 1 μM CASE004 in $1\times$ CS buffer for 200 s.
9. Remove unbound CASE004 and record baselines by dipping the sensor tips in $1\times$ CS buffer for 30 s.
10. Load RNA onto the tips by dipping them in the diluted RNA samples for 300 s.
11. Transfer the tips to $1\times$ CS buffer with appropriate target and record the release of RNA for 300 s.

12. Regenerate the biosensor tips by alternately dipping three times in regeneration buffer for 5 s and $1\times$ CS buffer for 5 s.
13. Repeat the loading of RNA step as before and measure dissociation in several dilutions of target compound.

4 Notes

1. During the reaction a white precipitate will form due to inorganic pyrophosphate generated as side product of RNA chain elongation. Some kits use inorganic pyrophosphatase (PPase) that catalyzes the hydrolysis of inorganic pyrophosphate to orthophosphate to prevent any inhibitory effects of pyrophosphate buildup. However, we have not had issues using kits without this addition.
2. A typical 30 min T7 transcription reaction will yield 4.5 nmol or 150 μg RNA per mL. The loss during removal of precipitate is less than 5%.
3. Be careful not to over-dry the pellet. If the pellet is too dry it will be difficult to redissolve. If the pellet is hard to dissolve, the RNA can be heated to 55–60 $^{\circ}\text{C}$ for 10 min to aid its dissolution.
4. It is very important to ascertain the fidelity of chosen ligands with the CS buffer. The presence of specific metals or surfactants or the pH of the buffer may precipitate ligand molecules or otherwise render selection of aptamers impossible.
5. Most thermocyclers allow the running protocol to be paused. This way samples can be drawn in the late extension phase without interfering with the next PCR cycle.
6. Subsequent rounds of RNA capture SELEX require less RNA (Subheading 3.1) and less bead preparation (Subheading 3.2). For round two and forward we usually make a total of 250 μL of T7 transcription reaction (a four-fold reduction) and load the RNA on to 10^8 beads (tenfold reduction).
7. Selected RNA pools from multiple rounds of Capture SELEX can also be screened as a way of evaluating the progress of the selection.
8. Alternatively, the RNA concentration can be estimated with the Bioanalyzer's built-in software; however, we find the Qubit to be much more reliable in measuring RNA concentrations.
9. CASE001 has an average molecular mass of 36450 Da. Thus, 200 μL of a 1 μM solution contains 7.3 μg RNA.

References

1. Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249:505–510
2. Ellington AD, Szostak JW (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature* 346:818–822
3. Lynch SA, Desai SK, Sajja HK, Gallivan JP (2007) A high-throughput screen for synthetic riboswitches reveals mechanistic insights into their function. *Chem Biol* 14:173–184
4. Lynch SA, Topp S, Gallivan JP (2009) High-throughput screens to discover synthetic riboswitches. *Methods Mol Biol* 540:321–333
5. Link KH, Breaker RR (2009) Engineering ligand-responsive gene-control elements: lessons learned from natural riboswitches. *Gene Ther* 16:1189–1201
6. Dietrich JA, McKee AE, Keasling JD (2010) High-throughput metabolic engineering: advances in small-molecule screening and selection. *Annu Rev Biochem* 79:563–590
7. Yang J, Seo SW, Jang S et al (2013) Synthetic RNA devices to expedite the evolution of metabolite-producing microbes. *Nat Commun* 4:1413
8. Nutiu R, Li Y (2005) In vitro selection of structure-switching signaling aptamers. *Angew Chem Int Ed Engl* 44:1061–1065
9. Vallée-Bélisle A, Plaxco KW (2010) Structure-switching biosensors: inspired by nature. *Curr Opin Struct Biol* 20:518–526
10. Suess B, Hanson S, Berens C et al (2003) Conditional gene expression by controlling translation with tetracycline-binding aptamers. *Nucleic Acids Res* 31:1853–1858
11. Muranaka N, Abe K, Yokobayashi Y (2009) Mechanism-guided library design and dual genetic selection of synthetic OFF riboswitches. *Chembiochem* 10:2375–2381
12. Topp S, Gallivan JP (2008) Random walks to synthetic riboswitches—a high-throughput selection based on cell motility. *Chembiochem* 9:210–213
13. Stoltenburg R, Nikolaus N, Strehlitz B (2012) Capture-SELEX: selection of DNA aptamers for aminoglycoside antibiotics. *J Anal Methods Chem* 2012:415697
14. Lauridsen LH, Sommer MOA, Nielsen AT (2015) Development of aptamers for in vivo and in vitro biosensor applications. Dissertation, Technical University of Denmark

High-Throughput Microfluidics for the Screening of Yeast Libraries

Mingtao Huang, Haakan N. Joensson, and Jens Nielsen

Abstract

Cell factory development is critically important for efficient biological production of chemicals, biofuels, and pharmaceuticals. Many rounds of the Design–Build–Test–Learn cycles may be required before an engineered strain meeting specific metrics required for industrial application. The bioindustry prefer products in secreted form (secreted products or extracellular metabolites) as it can lower the cost of downstream processing, reduce metabolic burden to cell hosts, and allow necessary modification on the final products, such as biopharmaceuticals. Yet, products in secreted form result in the disconnection of phenotype from genotype, which may have limited throughput in the Test step for identification of desired variants from large libraries of mutant strains. In droplet microfluidic screening, single cells are encapsulated in individual droplet and enable high-throughput processing and sorting of single cells or clones. Encapsulation in droplets allows this technology to overcome the throughput limitations present in traditional methods for screening by extracellular phenotypes. In this chapter, we describe a protocol/guideline for high-throughput droplet microfluidics screening of yeast libraries for higher protein secretion. This protocol can be adapted to screening by a range of other extracellular products from yeast or other hosts.

Key words Droplet microfluidics, High-throughput screening, Yeast cell factories, Random mutagenesis, Protein secretion, Systems biology

1 Introduction

Due to the cost, environmental impact, limitation in raw materials, product quality, and product quantity, there is an increasing interest in the development of cell factories for efficient production of chemicals, biofuels, pharmaceuticals as complementary to or substitute for chemical synthesis and natural extraction. Thanks to dramatic developments in systems biology and synthetic biology in past decades, more and more knowledge and tools are accessible to guide and implement host cells engineering [1–3]. Yet, due to the complexity of biological systems, many Design–Build–Test–Learn cycles may be required before an adequately engineered strain for industrial applications is obtained [4]. Large number of cell factory variants containing desired phenotype, ranging from

thousands to millions, can be built. However sensitive, convenient, and rapid methods to examine variants in the Test step are still needed.

Microtiter plate-based screening methods are widely used for screening of cell factories producing extracellular products [5–7]. This technology allows scientists to test hundreds to thousands of samples concurrently using limited amounts of reagents and consuming little space. Yet microtiter plate screening is laborious, low efficient, and time-consuming when it is applied to libraries of large numbers of variants ($>10^5$). High-throughput screening methods are required for evaluation these populations. Fluorescence activated cell sorting (FACS) is a mature technique providing high-throughput screening [8]. Up to 70,000 cells per second can be screened through FACS [9]. It has been successful utilized for isolation of improved astaxanthin production variants from a *Xanthophyllomyces dendrorhous* mutant library [10]. However, as a continuous-cell suspension flow microfluidic system, FACS is a cell-based screening method; it can only detect signals based on cytosolic products [11] or membrane attached products [12], hence limiting its applicability to screening for cells with extracellular phenotypic characterizations.

Droplet microfluidics, a technique where cells are encapsulated in picoliter droplets which can be sorted at high-throughput, represents a novel approach in the field of screening developed in the last decade [13, 14]. Compartmentalization in the picoliter droplet connects the cell with its immediate micro-environment. Droplet microfluidics has demonstrated its power in screening extra metabolites and secreted products [15–17]. Microfluidic devices, fabricated using methods developed in the semiconductor industry are used to generate and process picoliter droplets, greatly improving throughput and minimizing consumables usage. The rapid manufacturing processes used allow customized design of microfluidic for each different experiment. Here, we describe the experimental steps for production of microfluidic devices for droplet microfluidics and their usage in picoliter droplet microfluidic screening of yeast libraries for higher protein secretion as an example (Fig. 1). The protocol described here is not limited in screening yeast cells for amylase production, but can also be adapted for application on other extracellular products from yeast or other hosts.

2 Materials

2.1 Microfluidic Devices Assembly

1. Computer-aided design software (e.g. AutoCAD).
2. SU-8 photoresist (MicroChem).
3. Four-inch silicon wafer.

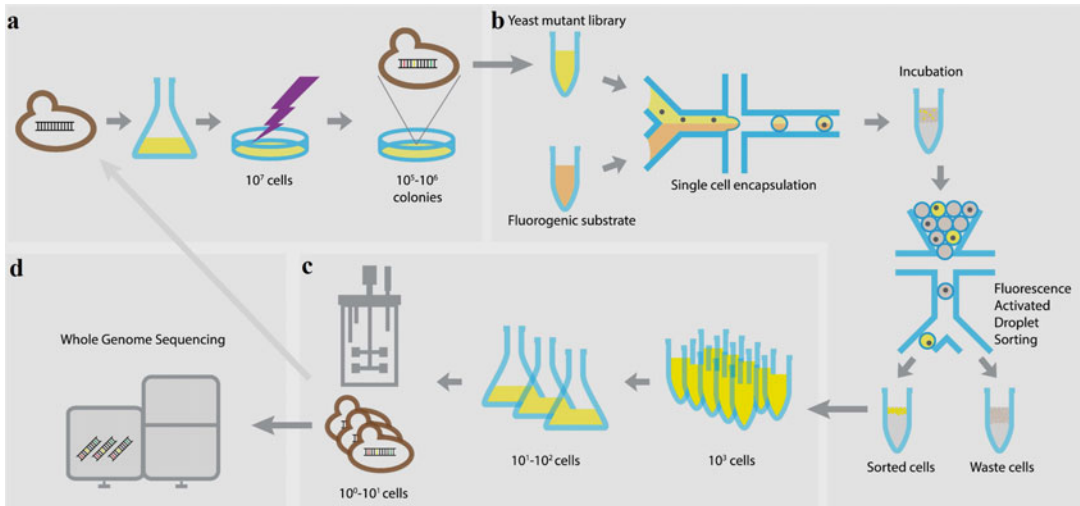


Fig. 1 The schematic workflow shows high-throughput microfluidics for screening of yeast libraries for higher α -amylase secretion. (a) Yeast mutant libraries are generated by UV mutagenesis. (b) Single cells are encapsulated in droplets together with fluorogenic substrate and sorted by the microfluidic screening chip. (c) Sorted cells are validated for α -amylase secretion and the best ones are continued to generate new libraries for accumulation of more benefit mutations or (d) analyzed for mutations associated with enhanced secretion via whole-genome sequencing. Adapted from ref. 16

4. Spin coater (e.g. Karl Suss Delta 20).
5. UV Photolithography Mask Aligner (Karl Süss).
6. $\geq 20,000$ DPI printed photolithography mask.
7. SU-8 developer: propylene glycol monomethyl ether acetate (PGMEA), (mrDEV 600, Microresist GmbH).
8. Polydimethylsiloxane (PDMS). Sylgard 184 (Dow Chemical).
9. $1'' \times 3''$ Glass microscope slide.
10. 0.75 mm and 2 mm Biopsy punch (Harris Uni-Core).
11. Bench-top Oxygen Plasma (FemtoScience).
12. Aquapel solution (PPG Industries).
13. Ultrasonic bath (Cole Parmer).
14. Low melting point solder (Indium Corp).
15. PCB pin header (Camdenboss).

2.2 Droplet Microfluidic System Operation

1. Inverted microscope (IX51, Olympus).
2. 491 nm, 50 mW laser (Cobolt AB).
3. Optical components (mirrors, dichroic mirrors, optical filters) (Semrock, Thorlabs).
4. Polyether ether ketone (PEEK) tubing $1/32''$ OD, $0.008''$ ID (Zeus Inc).
5. $1/32''$ fittings (IDEX H&S).

6. Syringe pumps (Harvard Apparatus and neMESYS, Cetoni GmbH).
7. BODIPY-starch (Invitrogen, EnzChek).
8. High-voltage amplifier unit (TREK Inc).
9. Data acquisition card with FPGA (NI PCIe-7841R, National Instruments).
10. Vibra-Cell ultrasonic processor (Sonics & Materials Inc).
11. Novec HFE-7500 oil (3 M).
12. Fluorosurfactant EA (RainDance Technologies).
13. Photomultiplier tube module (Hamamatsu).
14. Emulsion destabilizer, 1H,1H,2H,2H-Perfluoro-1-octanol (Sigma-Aldrich).

2.3 Yeast Library Preparation

1. α -amylase secretion strain: *Saccharomyces cerevisiae* CEN.PK 530.1C with *pAlphaAmyCPOT* [18].
2. YPD medium: 10 g/l yeast extract, 20 g/l peptone, and 20 g/l glucose.
3. SD-2 \times SCAA medium [16]: 20 g/l glucose, 6.9 g/l yeast nitrogen base without amino acids, 190 mg/l Arg, 400 mg/l Asp, 1260 mg/l Glu, 130 mg/l Gly, 140 mg/l His, 290 mg/l Ile, 400 mg/l Leu, 440 mg/l Lys, 108 mg/l Met, 200 mg/l Phe, 220 mg/l Thr, 40 mg/l Trp, 52 mg/l Tyr, 380 mg/l Val, 1 g/l BSA, 5.4 g/l Na₂HPO₄, and 8.56 g/l NaH₂PO₄·H₂O (pH = 6.0 by NaOH). For bioreactor fermentations, 5.4 g/l Na₂HPO₄ and 8.56 g/l NaH₂PO₄·H₂O in the SD-2 \times SCAA were replaced by 2 g/l KH₂PO₄ (pH = 6.0 by NaOH) (*see Note 1*).
4. Starch agar plate: 0.04 g/l glucose, 10 g/l starch, 6.9 g/l yeast nitrogen base without amino acids, and 20 g/l agar.
5. 9 cm Petri dishes and 14 cm Petri dishes (*see Note 2*).
6. UV cross-linker (Topac Inc.) with 40 W lamp at 254 nm.

2.4 Evaluation and Analysis of Sorted Cells

1. α -amylase assay (K-CERA) kit (Megazyme).
2. α -amylase power from *Aspergillus oryzae* (Sigma).
3. Block heater (SBH130DC, Stuart).
4. Spectrophotometer (GENESYS 20, Thermo Scientific).
5. NuPAGE LDS Sample Buffer.
6. NuPAGE Sample Reducing Agent.
7. Gradient (4–20%) precast polyacrylamide gel (Bio-Rad).
8. Coomassie blue.
9. Mini-PROTEAN Tetra System (Bio-Rad).
10. Blood & Cell Culture DNA kit (Qiagen).

3 Methods

3.1 Manufacturing of PDMS Master Mold

1. Design or adapt previous design for microfluidic droplet generation circuit and droplet sorting circuit in CAD software.
2. Print high-resolution photomask based on CAD circuit design.
3. Distribute appropriate thickness SU-8 on a 4-in. silicon wafer using a spin coater.
4. Bake the SU-8-coated silicon wafer at 65 °C and 95 °C according to SU-8 manufacturer's instructions for the desired thickness.
5. Expose SU-8-coated silicon wafer with UV light through photolithography mask in a mask aligner. UV dose is dependent on SU-8 thickness.
6. Bake the wafer at 65 °C and 95 °C according to SU-8 manufacturer's instructions for the desired thickness.
7. Immerse the Si-wafer in SU-8 developer to dissolve unexposed SU-8 for a length of time depending on SU-8 thickness.
8. Clean the patterned wafer by immersing it in isopropanol.
9. Blow dry the patterned Si-wafer master mold using compressed N₂ gas.
10. Hard bake the patterned Si-wafer at 150 °C for 30 min to anneal any surface cracks in the patterned SU-8.

3.2 Fabrication of PDMS Glass Droplet Microfluidics Devices

1. Mix Sylgard 184 PDMS base with curing agent at a 10:1 ratio.
2. Pour ca 90 g of the PDMS mixture onto the master mold in a 5" petri dish.
3. Degass the PDMS on the master mold in a degassing chamber to remove air. Bubbles from the patterned Si-wafer surface.
4. Bake the PDMS covered master for >2 h at 65 °C.
5. Gently cut out and peel off the PDMS slab from the master mold using a scalpel.
6. Cut inlet and outlet holes through PDMS slab for channel access using biopsy punches. Use 0.75 mm diameter punches for liquid channels and 2 mm diameter punches for electrode channels. Cut from the channel side of the slab.
7. Clean the PDMS slab from debris by repeatedly rubbing it with scotch tape.
8. Treat the PDMS slab with oxygen plasma to activate its surface. Immediately following plasma treatment, press a glass microscope slide onto the channel side of the PDMS (*see Note 3*).
9. Inject the aquapel solution to the chip for fluorophilic surface treatment, immediately flush with filtered nitrogen gas.

10. Seal all liquid channels by covering the microfluidics chip with a length of scotch tape.
11. For fabrication of electrodes on the sorting chip, heat the chip and low melting point solder to 100 °C on a hot plate. Inject the liquid solder into designed electrode channels using a syringe and 2 mm diameter short blunt needle. Insert PCB pin header to electrode channels for off-chip voltage connection. PCB pin header will be fixed in electrode channel inlets after liquid solder solidification. Control electrode integrity by determining the Ohmic resistance between electrode inlet and outlet channel.

3.3 Yeast UV-Mutagenesis Library Generation

1. Strike out the α -amylase secretion yeast strain from glycerol stock in -80 °C freezer on a YPD plate; incubate the plate at 30 °C for 24–48 h for colony development (*see Note 4*).
2. Pick one colony from the fresh YPD plate to 3 ml YPD medium; incubation overnight at 30 °C with 200 rpm shaking.
3. Harvest yeast cells and wash with distilled water twice; resuspend the cell pellet in distilled water to $OD_{600} = 2$ (*see Note 5*).
4. Spread 300 μ l cell suspensions on starch agar plate.
5. Put the starch agar plate into the UV cross-linker; expose cells under UV light for 2–8 s (corresponding to 4–16 mJ/cm²) (*see Note 6*).
6. Label plates with UV treatment intensity; incubate UV irradiation plates at 30 °C in dark until colonies formed (*see Note 7*).
7. Store plates at 4 °C before proceeding to sorting.

3.4 Microfluidic Experiment Preparation

1. Place the microfluidic device on an adjustable microscope xy-table.
2. Connect syringes to the microfluidic device with 1/32" PEEK tubing and fittings.
3. Mount cell suspension syringe on a Harvard Apparatus syringe pump and control all additional syringes on neMESYS syringe pumps.
4. Focus the microscope on the main channel of the microfluidic device using 10 \times objective. Direct a 491 nm laser to the microscope objective to illuminate droplets and induce fluorescence. The fluorescence intensity of each single microdroplet is detected at 525 ± 20 nm using suitable band pass filters and a photomultiplier tube module connected to the data acquisition card.
5. Connect the input port of the high-voltage amplifier unit to an analogue output on the data acquisition card and the output to the electrodes of the microfluidic device.

6. Create an electric field on the microfluidic device by 1000-fold amplification of a computer generated ca $1V_{p-p}$ 30 kHz square wave signal.

3.5 Encapsulation of Yeast Library in Droplets

1. Collect yeast mutant library from the starch plate with 1 ml SD-2×SCAA medium.
2. Ultrasonicate cell suspension at 40 W for 10 s three times with the Vibra-Cell ultrasonic processor (*see Note 8*).
3. Wash cells three times with medium and resuspend in SD-2×SCAA medium (BSA increased to 5%) at around 1.5×10^6 cells/ml; put cell suspension on ice (*see Note 9*).
4. Transfer cell suspension to a syringe and 200 $\mu\text{g}/\text{ml}$ BODIPY-starch substrate to another syringe, and connect syringes to aqueous inlets of the generation chip.
5. Simultaneously inject cell suspension and BODIPY-starch substrate into the droplet generation circuit and mix them in a T-connector at 1:1 ratio. Run the generation chip at a total aqueous flow rate of 200 $\mu\text{l}/\text{h}$ and the oil (HFE-7500 with 1% EA surfactant, a polyethylene glycol perfluoropolyether amphiphilic block copolymer) flow rate of 1000 $\mu\text{l}/\text{h}$. Yeast cells are encapsulated in the droplets. This can produce 20 pl droplets at a rate of approximately 3000 droplets per second with a cell encapsulation ratio of yeast cell to droplet of approximately 0.4 (*see Note 10*).
6. Collect encapsulated cell emulsion in a 1 ml plastic syringe.

3.6 Sorting of Yeast Cells Encapsulated in Droplets

1. Incubate the emulsion in the syringe at room temperature for 3 h.
2. Connect the syringe containing emulsion and the syringe containing HFE-7500 oil with 1% EA surfactant to emulsion inlet and oil inlet of the sorting chip, respectively.
3. Connect the input port of the high-voltage amplifier unit to an analogue output on the data acquisition card and the output to the electrodes of the microfluidic device.
4. Create an electric field on the microfluidic device by 1000-fold amplification of a computer generated ca $1V_{p-p}$ 30 kHz square wave signal.
5. Inject the emulsion at a flow rate of 30 $\mu\text{l}/\text{h}$ and the oil at a flow rate of 300 $\mu\text{l}/\text{h}$.
6. Acquire the fluorescence intensity from single droplets when it passes the sorting junction using PMT connected to FPGA data acquisition card. Display droplet fluorescence using Lab-View software.
7. Set a sorting threshold based on the fluorescence distribution of the droplet population to the FPGA on the data acquisition

card. For droplet signals above this threshold, apply an electric field on the microfluidic device by 1000-fold amplification of an FPGA-generated ca $1V_{p-p}$ 30 kHz square wave signal pulse of approximately 800 μ s. Droplets with fluorescence over the threshold will be directed to the collection syringe through the sorted channel (*see Note 11*).

3.7 Extraction of Yeast Cells After Sorting

1. Remove excessive oil from the collection syringe.
2. Add 5 μ l emulsion destabilizer to sorted droplets and vortex gently.
3. Add 300 μ l fresh medium to the broken emulsion.
4. Take the upper phase containing yeast cells and spread on starch plates.
5. Incubate starch plates at 30 °C until colonies formed.

3.8 Evaluation of Sorted Cells

1. Pick mutant colonies from starch plates to 2.5 ml SD-2 \times SCAA medium in 14 ml falcon tubes; inoculate at 30 °C with 200 rpm shaking for 96 h.
2. Measure the optical density of culture at 600 nm.
3. Take 500 μ l cultured medium to a 1.5 ml EP tube and centrifuge at 14,000 $\times g$ for 5 min.
4. Transfer the supernatant to a new 1.5 ml EP tube and determine amylase activity with the amylase assay kit.
5. Calculate the amylase titer and amylase yield of mutant strains.
6. Analyze the amount of amylase in the supernatant with SDS-PAGE.
7. Test some of best performing mutant strains with shake flask or bioreactor fermentation.
8. The best ones can be used as starting strains for a new round of UV mutation and screening (*see Note 12*).

3.9 Analysis of Mutations of Sorted Cells

1. Inoculate selected sorted cells to YPD medium at 30 °C with 200 rpm shaking.
2. Harvest cells at late exponential phase to stationary phase, use 15–20 OD cells for genomic DNA extraction with the Blood & Cell Culture DNA kit (*see Note 13*).
3. Prepare DNA samples using the Illumina DNA TruSeq protocol.
4. Sequence DNA samples using the version 2 chemistry on an Illumina MiSeq, paired-end 500 cycles (2 \times 250 bp).
5. Map reads to the reference genome of *S. cerevisiae* CEN.PK 113–7D using MosaikAligner version 2.1.32. Postprocess the alignments to realign potential indels and remove likely PCR

duplicates using GATK 2.3.9 [19] and Picard tools 1.100 (picard.sourceforge.net), respectively.

6. Detect single nucleotide variants and small indels with GATK UnifiedGenotyper and annotate with SnpEff 3.4. Detect large-scale chromosome duplications by plotting the mapped coverage over all chromosomes. Detect smaller structural variants with SVseq2 [20].
7. Inspect all detected variants manually in a genome browser to detect and discard obvious false positives.
8. Analyze gene mutations based on the information and tools from the Saccharomyces Genome Database.

4 Notes

1. Dissolve sodium phosphate and YNB in two different beakers, mix together until both of them dissolve completely in distilled water to avoid precipitation in the solution.
2. Use 14 cm petri dishes for library preparation. Bigger petri dishes can hold more cell colonies.
3. To ensure bonding the PDMS/glass chip can at this point be baked at 95 °C for >2 h.
4. A fresh strain from the glycerol stock is preferred. Don't prepare the library with cells from plates in long-term store at 4 °C to avoid starting with degenerate strains.
5. Cells should be harvested at the exponential phase to early stationary phase. Cells from death phase may lose robustness in recovery on plates and be less sensitive to UV light, which may reduce mutation efficiency.
6. Cell death rate between 85% and 95% is recommended. From our experience, UV dose between 7 and 12 mJ/cm² results in good libraries. Almost all cell die when using UV dose over 16 mJ/cm².
7. It is important to incubate strains in dark after UV irradiation in order to avoid photoreactivation repair.
8. This ultrasonic step is able to loose clustered cells to single cells, which increase the accuracy in screening. If more cells are encapsulated into one droplet, there will be more amylase secreted in the droplet; and it could lead to a mis-sorting of droplets.
9. A suitable cell density is important for efficient cell encapsulation. Too high cell density leads to a greater proportion of encapsulation of more cells into one droplet. Too low cell density results in too many empty droplets following

encapsulation, limiting throughput. The cell density of 1.5×10^6 cells/ml results in an appropriate ratio of yeast cell to droplet at 0.4.

10. Adjust pulse length according to exact droplet re-injection rate to ensure that only a single droplet is diverted per pulse.
11. Check for a stable droplet generation rate by assaying droplets for background fluorescence immediately following droplet generation. Unstable droplet generation rates will result in a greater variation in droplet size and subsequently a lower assay resolution.
12. Besides UV irradiation, other mutagenesis strategies (different irradiation energies, mutagenic chemicals, etc.) can apply for generation of libraries in order to achieve large diversity mutations.
13. The mother strain, which is used for library generation, should be sequenced together with the mutant strains. Nonexistent mutations may be detected due to potential differences between the mother strain and online reference strain, if just mapping sequencing results of mutant strains to the online reference genome without comparing with the sequencing result of the mother strain.

References

1. de Jong B, Siewers V, Nielsen J (2012) Systems biology of yeast: enabling technology for development of cell factories for production of advanced biofuels. *Curr Opin Biotechnol* 23(4):624–630
2. Khalil AS, Collins JJ (2010) Synthetic biology: applications come of age. *Nat Rev Genet* 11(5):367–379
3. Jinek M et al (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337(6096):816–821
4. Nielsen J, Keasling Jay D (2016) Engineering cellular metabolism. *Cell* 164(6):1185–1197
5. Kitagawa T et al (2011) Identification of genes that enhance cellulase protein production in yeast. *J Biotechnol* 151(2):194–203
6. Babiskin AH, Smolke CD (2011) A synthetic library of RNA control modules for predictable tuning of gene expression in yeast. *Mol Syst Biol* 7:471
7. Shi SB, Ji HC, Siewers V, Nielsen J (2016) Improved production of fatty acids by *Saccharomyces cerevisiae* through screening a cDNA library from the oleaginous yeast *Yarrowia lipolytica*. *FEMS Yeast Res* 16(1):fov108
8. Herzenberg LA et al (2002) The history and future of the fluorescence activated cell sorter and flow cytometry: a view from Stanford. *Clin Chem* 48(10):1819–1827
9. Eisenstein M (2006) Cell sorting: divide and conquer. *Nature* 441(7097):1179–1185
10. Ukibe K, Katsuragi T, Tani Y, Takagi H (2008) Efficient screening for astaxanthin-overproducing mutants of the yeast *Xanthophyllomyces dendrorhous* by flow cytometry. *FEMS Microbiol Lett* 286(2):241–248
11. Carqueijeiro I et al (2016) Isolation of cells specialized in anticancer alkaloid metabolism by fluorescence-activated cell sorting. *Plant Physiol* 171(4):2371–2378
12. Wentz AE, Shusta EV (2007) A novel high-throughput screen reveals yeast genes that increase secretion of heterologous proteins. *Appl Environ Microbiol* 73(4):1189–1198
13. Baret JC et al (2009) Fluorescence-activated droplet sorting (FADS): efficient microfluidic cell sorting based on enzymatic activity. *Lab Chip* 9(13):1850–1858
14. Brouzes E et al (2009) Droplet microfluidic technology for single-cell high-throughput screening. *Proc Natl Acad Sci U S A* 106(34):14195–14200
15. Wang BL et al (2014) Microfluidic high-throughput culturing of single cells for selection based on extracellular metabolite

- production or consumption. *Nat Biotechnol* 32(5):473–478
16. Huang M et al (2015) Microfluidic screening and whole-genome sequencing identifies mutations associated with improved protein secretion by yeast. *Proc Natl Acad Sci U S A* 112(34):E4689–E4696
 17. Sjostrom SL et al (2014) High-throughput screening for industrial enzyme production hosts by droplet microfluidics. *Lab Chip* 14(4):806–813
 18. Liu Z, Tyo KEJ, Martínez JL, Petranovic D, Nielsen J (2012) Different expression systems for production of recombinant proteins in *Saccharomyces cerevisiae*. *Biotechnol Bioeng* 109(5):1259–1268
 19. McKenna A et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297–1303
 20. Zhang J, Wang JY, Wu YF (2012) An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. *BMC Bioinformatics* 13:S6

Growth-Coupled Carotenoids Production Using Adaptive Laboratory Evolution

Luis H. Reyes and Katy C. Kao

Abstract

Adaptive laboratory evolution is a powerful technique for strain development. However, the target phenotypes using this strategy have been limited by the required coupling of the phenotype-of-interest with fitness or survival, and thus adaptive evolution is generally not used to improve product formation. If the desired product confers a benefit to the host, then adaptive evolution can be an effective approach to improve host productivity. In this book chapter, we describe an effective adaptive laboratory evolution strategy for improving product formation of carotenoids, a class of compounds with antioxidant potential, in the yeast *Saccharomyces cerevisiae*.

Key words Adaptive laboratory evolution, Yeast, Carotenoids, Antioxidants, Strain improvement

1 Introduction

Adaptive laboratory evolution (ALE) is recognized as a powerful method for improvement of several phenotypes of industrial strains without requiring extensive knowledge of underlying molecular basis of these phenotypes [1–3]. The main caveat with using ALE is that it requires the desired trait to be directly linked with a growth benefit to the microorganism. In ALE, adaptive mutants (strains with improved fitness) arise spontaneously and expand in a population in environments with defined selective pressures. The design of suitable selective pressures is the foundation for successful ALEs. There is a saying in the ALE field that, in colloquial terms, “*you get what you select for*”, which captures the importance of choosing the proper selective pressure for directing ALE.

Engineered strains with increased productivity of a desired compound usually exhibit reduced fitness. During whole-cell biocatalysis using engineered strains, the metabolic intermediates necessary for cell growth and maintenance are depleted, being redirected for biosynthesis of desired compounds. This drainage of energy and biomass precursors, including nucleotides and amino

acids, imposes a metabolic load to the cell, generally resulting in a reduction of biomass-related proteins [4, 5]. This negative impact on cell growth in the production of synthetic and secondary metabolites hinders the use of ALE for improvement of their production. The key in designing an ALE experiment for increased synthetic metabolite production lies in how much its production benefits cell growth. However, since the production of desired compounds confers a fitness reduction, it poses an evolutionary disadvantage and a barrier for the use of ALE for strain development. This barrier may be overcome if an appropriate selective pressure can be designed to counteract the metabolic burden in favor of growth advantage, making ALE a potentially powerful tool to develop better producers.

In this chapter we describe a methodology designed for the improvement of carotenoids production in *Saccharomyces cerevisiae*. Carotenoids, also known as tetraterpenoids, are highly hydrophobic organic pigments that are naturally produced by several organisms from plants to bacteria. These compounds are of particular interest due to their antioxidant properties and pigmentation, with uses in nutraceuticals, cosmetics, and the food industry. The biosynthesis of all carotenoids uses isopentenyl diphosphate (IPP) as precursor; the biosynthetic routes are described elsewhere [6]. Extensive metabolic engineering efforts have been applied to develop *S. cerevisiae* strains for the heterologous production of various carotenoids [6–8]. IPP is also a precursor of other isoprenoids, either via MVA (mevalonic acid) pathway in eukaryotes, or the DXP (1-deoxy-D-xylulose 5-phosphate; also known as methylerythritol phosphate) pathway found in several bacteria. In *S. cerevisiae*, the MVA pathway is predominantly used for ergosterol (provitamin D2) biosynthesis, which is an essential component of the yeast membrane and provides membrane permeability and fluidity [9]. Thus, the heterologous production of carotenoids directly affects growth of the fungal host cell due to the redirected IPP flux towards carotenoids production instead of ergosterol biosynthesis. Indeed, prolonged propagation of engineered carotenogenic yeast in normal culture conditions leads to no improvement in carotenoids productivity and in some cases a net loss of productivity (data not shown). To develop an ALE strategy for improving carotenoids production in yeast, we theorized that oxidative stress could be used as a driving force for the directed evolution of *S. cerevisiae* for enhanced carotenoids production. Upon first identifying the exact condition that allows the production of carotenoids to confer protection against oxidative stress (hydrogen peroxide challenge), our prior work successfully improved heterologous carotenoids production in *S. cerevisiae* [1].

While the work described here is specific for carotenoids production in *Saccharomyces cerevisiae*, this approach can potentially be applied to improve production of other compounds with

antioxidant potential. This method has been recently used to increase 3'-hydroxygenistein production in *Pichia pastoris* and to identify overproducers of lycopene in *Escherichia coli* [10, 11].

2 Materials

2.1 Biological and Chemical Materials

1. Target strain previously engineered for the production of the desired compound with antioxidant properties to be used for adaptive laboratory evolution.
2. Growth media for the course of the evolutionary experiment. Defined media (Yeast Nitrogen Base YNB) is preferred in most cases since it allows better control of nutrients and other conditions. However rich media (Yeast-Peptone-Dextrose YPD) may be used as necessary.
3. Inhibitor stocks. The inhibitor described in this chapter is hydrogen peroxide. Hydrogen peroxide (stock concentration 35% w/w) was used in order to apply oxidative stress to the microbial cultures, although the chosen inhibitor is highly dependent of the endogenous resistance of the microorganism and the requirements of the evolution experiment. Inhibitor stocks should be filter-sterilized.
4. 50% (v/v) sterile glycerol for cryogenic storage of evolving populations and isolated mutants.
5. 425–600 μm acid-washed glass beads (Sigma) for cell disruption.
6. Dodecane anhydrous $\geq 99\%$ (Sigma) for the extraction of the hydrophobic β -carotene and other carotenoids.

2.2 Equipment

1. Bench-top centrifuges.
2. Isothermal incubators and shakers.
3. Spectrophotometer.
4. Microplate reader (e.g., TECAN Infinite[®] M200) for high-throughput analysis, with capabilities of survey scans from OD₃₅₀ to OD₅₅₀.
5. Optical microscope (e.g., Zeiss Axio Scope.A1) for visual inspection of morphology. For yeast, a 40 \times phase contrast objective (for a total of 400 \times magnification combined with the 10 \times magnification of the eye piece) is sufficient. For bacteria, a 100 \times phase contract objective is needed.
6. Culture tubes and/or Erlenmeyer flasks.
7. Disruptor Genie[®] Cell Disruptor (Scientific Industries).

Table 1
List of primers used in this study

| Name | Sequence | Notes |
|----------|---|---|
| NEO_f | 5' - TTA AAA AAA TCC TTC TCT TGT CTC ATG CCA ATA AGA TCA ATC AGC TCA GCT TCA CAA ATG <u>CGG ATC</u> <u>CCC GGG TTA ATT AA</u> - 3' | The first 60 bp of each primer correspond to homologous regions in the yeast genome, flanking the <i>CTTI</i> gene. The following 20 bp (underlined) are complementary to the NEO cassette. |
| NEO_r | 5' - TAT AAT TAC GAA TAA TTA TGA ATA AAT AGT GCT GCC TTA ATT GGC ACT TGC AAT GGA CCA <u>GAA</u> <u>TTC GAG CTC GTT TAA AC</u> - 3' | |
| KO.Ver_f | 5' - ATT CGA CGT AGC CTG GAC AC - 3' | Primers used for the verification via PCR of the <i>ctt1Δ</i> knockout strains |
| KO.Ver_r | 5' - TAA TCG TTG AGT TCA TGC CG - 3' | |

2.3 Molecular Biology

1. Strain genotype: *S. cerevisiae* strain GSY1136 (Mat α , *ura3-52*, *gal +* in S288c background, YBR209W::*Act1p-GFP-Act1t-URA3*) [12].
2. Primers used are listed in Table 1.
3. Plasmids used: YIplac211YB/I/E* shuttle vector [6] for the production of carotenoids in yeast, and pFA6a-kanMX6 (Addgene plasmid # 39296 [13]) for the deletion of the *CTTI* gene.

3 Methods

The most commonly used strategy for adaptive laboratory evolution is continuous exposure to the chosen stressor, either via *serial batch transfers* or *continuous cultures* [14]. However, this may not always be the best strategy, especially when the production of the desired product poses a metabolic burden (*see Note 1*). In the following section we introduce a modified strategy for adaptive laboratory evolution experiments using a *Periodic Challenge/Recovery Scheme*. This modified method can be implemented when the growth benefit of the produced antioxidant compound is not sufficiently large to compensate for the metabolic load imposed on the microorganism; such is the case for carotenoids production in yeast. Below, we describe in detail the necessary steps for using this methodology in order to increase the carotenoids production in *S. cerevisiae*. Finally, some notes concerning contamination and instances where the described methodology may not be suitable are including in Subheading 4.

3.1 Construction of Carotenoids-Producing Strains

The methods described below correspond to the modification of the *S. cerevisiae* strain GSY1136, however they can be modified for the production of carotenoids in other yeast strains.

1. Since the GSY1136 strain used in this work contains the URA3 marker, it was necessary to excise it in order to create a uracil auxotrophic strain. We recommend using the anti-metabolite 5-fluoroorotic acid (5-FOA) as counter-selection to select for the loss of the URA3 marker, as described elsewhere [15]. This procedure was used to generate the uracil auxotrophic strain GSY1136 *ura3Δ*.
2. Digest the plasmid YIplac211YB/I/E* using the restriction endonuclease *StuI* and purify the linearized plasmid.
3. Transform 1 μg of linearized plasmid into GSY1136 *ura3Δ* using the lithium-acetate procedure or via electroporation to generate the carotenogenic yeast strain GSY1136 YIplac211YB/I/E*.
4. Plate the transformed cells onto solid YNB without amino acids (or just without uracil supplementation) and incubate at 30 °C for 3 days.
5. The correct constructs can be visually verified since the produced carotenoids have a characteristic red color.

3.2 Construction of Catalase-Deficient Strains

The rationale behind using ALE to increase carotenoids production is based on the antioxidant properties of these compounds. *S. cerevisiae* possesses two catalases (CTT1 and CAT1) [16]. It has been demonstrated that the cytosolic catalase CTT1 is essential for protecting the yeast cell against hydrogen peroxide stress [17]. Therefore, it is necessary to delete CTT1 to significantly reduce the native H₂O₂ detoxification mechanisms in yeast, in order to increase the antioxidant benefits of carotenoids. The outline for the deletion of CTT1 via homologous recombination in the yeast strain GSY1136 YIplac211YB/I/E*, is described below; however, alternative methods of disrupting the catalase gene in other yeast strains can also be used. For a different microbial platform, the primers should be modified accordingly.

1. Amplify the NEO cassette from the plasmid pFA6a-kanMX6 using the primers *NEO_f* and *NEO_r*.
2. Purify the amplicon and transform it into competent yeast cells using the lithium-acetate procedure or via electroporation.
3. Plate the cells onto solid media supplemented with geneticin (G418) at 200 μg/ml for selection.
4. Incubate the cells for 2 days or until colonies are observed.
5. Verify knockout strains (GSY1136 YIplac211YB/I/E* *ctt1Δ*) by PCR, using the primers *KO.Ver_f* and *KO.Ver_r*.

6. In our studies, we did not observed significant differences in growth rates between the parental and *ctt1Δ* strains in YNB + glucose or YPD, incubated at 30 °C.

3.3 Periodic Challenge/Recovery Scheme

Serial batch transfers have been routinely used to carry out adaptive evolution experiments [2, 18]. This method involves the serial transfer of evolving populations into media containing an inhibitor used as selective pressure, and the concentration of this inhibitor is generally increased periodically (ramped-up) (*see Note 2*). Initially we hypothesized that by using serial batch transfers, as depicted in Fig. 1a, to perform the evolution experiments, the antioxidant action of carotenoids will alleviate the oxidative stress caused by mild H₂O₂ concentrations close to the MIC₅₀ (defined as the minimal inhibitory concentration [MIC] of an inhibitor necessary to reduce growth of a target microorganism by 50%). However, we discovered that when cells were actively growing, the accumulation of carotenoids within the cell was not sufficient for alleviating H₂O₂ stress; and instead of the intended directed evolution towards increased carotenoids production, the productivity of the population actually decreased over time (data not shown). It is likely that in this case, since the production of the desired metabolite imposes a metabolic burden to the cell, the continuous exposure of a low level of oxidative stress was not sufficient for the benefit of carotenoids production to overcome the metabolic burden.

Preliminary experiments were carried out to determine the exact condition in which the production of carotenoids confers a benefit to the producing host. Since carotenoids are constitutively produced using the pTDH3 promoter [6], the maximum intracellular yield of the bio-product was expected to occur at the end of the exponential growth phase. We compared the cell viability of cell cultures from early-, mid-, and late-exponential phase exposed to high (lethal) concentrations of hydrogen peroxide in short periods of time, and a higher survival of yeast cells was observed in cells challenged during late-exponential growth.

Therefore, a new strategy using *Periodic Challenge/Recovery* was implemented. This modified methodology aims to fully exploit the antioxidant capabilities of carotenoids. Instead of continuous exposure of the cells in a stressful environment, as in usual evolutionary experiments, the populations were cultured in the absence of stressor until late-exponential/early stationary phase for two reasons: first to generate a larger amount of biomass and second to maximize the accumulation of carotenoids inside the cells. Periodic challenge of carotenoids-producing cells in the appropriate growth phase with high concentrations of hydrogen peroxide for short periods of time (shocking periods) (Fig. 1b) is then used to select for mutants with enhanced carotenoids production, as better producers are expected to exhibit increased survival. During the periodic challenge, ~95% of the population die, but cells producing

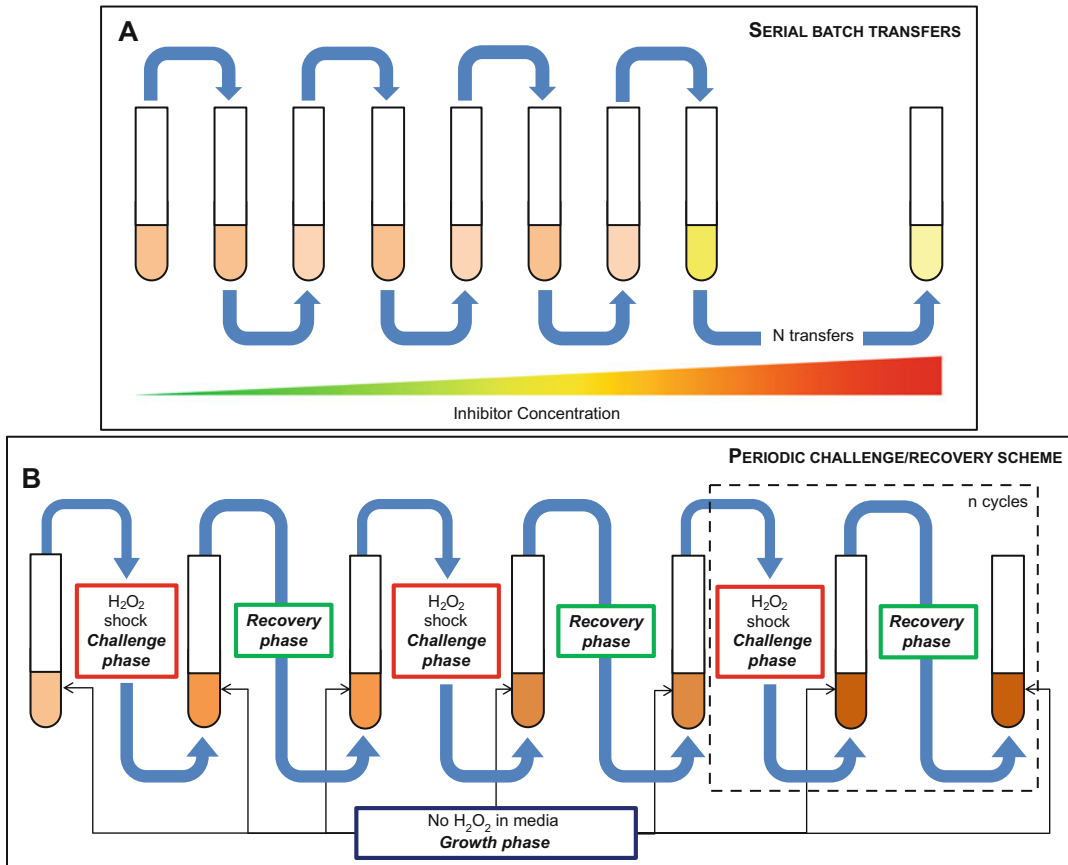


Fig. 1 (a) The *Serial Batch Transfer Method*, the most commonly used strategy for adaptive laboratory evolution where the microorganisms are continuously exposed to a chosen stressor. The selective pressure may be increased throughout the course of the adaptive laboratory evolution. (b) The modified method called *Periodic Challenge/Recovery Scheme*. This two-phase strategy uses a *challenge* phase using high concentrations of hydrogen peroxide to select for cells with larger antioxidants properties, and a *recovery* phase without the stressor to allow the cells more time to recover after peroxide shock

the highest amounts of carotenoids have better chance of survival. This *challenge* phase is followed by a *recovery* phase to allow the cells more time to recover from the hydrogen peroxide shock. This approach should significantly reduce the contribution of metabolic burden from the equation. Preliminary experiments should be carried out for the specific host producer strain, and the specific stressor used to determine the appropriate strategy to use, especially when to impose the challenge phase.

3.4 Determination of Sensitivity of Production Host to Oxidative Stress

Before starting any adaptive laboratory evolution experiment, it is indispensable to determine the inhibitory effect of the compound used to create the selective pressure. For this reason, it is crucial to determine the amount of H₂O₂ for the ALE. This amount is based on determining the effect of the selective pressure on cell viability; a

bottleneck of 5% survival is an arbitrary, but reasonable starting point. It is recommended to use a non-producing strain as control to identify if the effect of β -carotene on H_2O_2 alleviation is large enough to confer a fitness advantage to the producer strain.

1. Streak out the strain of interest onto appropriated agar plates. Incubate the plates at permissible temperature until colonies are visible.
2. Pick one colony for every set of biological replicates and inoculate 3 ml of liquid media (same as one to use in evolution experiments). Grow cells at the appropriated conditions to maximize carotenoids production.
3. Normalize cultures by cell density and transfer 500 μl of the normalized culture to microcentrifuge tubes.
4. Add various amounts of hydrogen peroxide to the samples, to get concentrations ranging from 0.5 M to 3.0 M. Mix samples by vortex for 3 s. The amount of hydrogen peroxide needed may vary depending of the strain being studied.
5. Allow the reaction to proceed for 30 min. It is recommended to keep the tubes lids open in a sterile environment (e.g., laminar flow hood) due to the continuous release of O_2 from the reaction.
6. Pellet cells in a bench-top centrifuge at $3000 \times g$ for 5 min.
7. Remove supernatant and resuspend cells in fresh liquid media.
8. Plate cells onto agar plates at several dilutions and determine cell viability. A concentration of hydrogen peroxide ensuring a cell survival around a target frequency (e.g., 5%) may be used as the starting point for the implementation of periodic challenge/recovery strategy.

3.5 Periodic Challenge/Recovery Experiments

1. Streak out strain of interest onto agar plates. Grow cells at permissible temperature until colonies are visible.
2. Pick one colony for every set of technical replicates and inoculate it into liquid media. Replicated cultures from independent colonies are necessary in case of events of genetic drift due to the large bottleneck that the periodic challenge/recovery experiments represent.
3. Make glycerol stocks of the overnight cultures to save the parental strain as control for subsequent analyses. This is especially important when identification of mutations associated with improved productivity is desired, since jackpot mutations may already be present in the inoculum.
4. Prepare X test tubes or Erlenmeyer flasks with γ ml of media.
 - (a) The number of replicates (X) depends on the number of biological and technical replicates (ideally a minimum of

three independent colonies as biological replicates with two technical replicates each).

- (b) The media volume (γ) depends on culture vessel used in the experiment. As a rule of thumb, the volume of the culture should be 20% of the total volume of the culture vessel to ensure adequate mixing and aeration.
5. Inoculate the cultures with the overnight culture using 1–2% (v/v) inoculum (e.g., for a 100 ml culture, use 1–2 ml of overnight culture).
6. Incubate the cultures at the appropriated conditions to maximize carotenoids production.
7. Once the samples reach the desired condition (late-exponential phase in our experience):
 - (a) Quantify the carotenoids concentration using the protocol described in Subheading 3.5.1.
 - (b) Prepare X sterile 1.5 ml microcentrifuge tubes for hydrogen peroxide shock.
 - (c) Prepare glycerol stocks of the evolving populations using 67% culture and 33% by volume 50% glycerol. Store at 80 °C for later analysis.
 - (d) Prepare X more tubes (or Erlenmeyer flasks) with γ ml of liquid media.
8. Add the amount of hydrogen peroxide determined from Subheading 3.3 to the samples. Mix samples by vortex for 3 s.
9. Allow the reaction to proceed for 30 min taking into account the recommendations from Subheading 3.3.
10. Pellet cells in a bench-top centrifuge at $3000 \times g$ for 5 min.
11. Remove supernatant and resuspend cells in fresh liquid media.
12. Inoculate X new cultures with 5% (v/v) inoculum from treated or untreated (control or during recovery phase) cells and cultivate until the cultures reach stationary phase.
13. Repeat **steps 6, 7 and 12** to allow the cells to recover after the hydrogen peroxide shock treatment (serial passage without a hydrogen peroxide challenge).
14. Repeat **steps 6–12** for a new round of peroxide challenge.
15. Repeat **steps 13 and 14** until the desired experimental phenotypic goals have been reached or for a set number of generations (*see* **Notes 3 and 4**).

3.5.1 Quantification of Carotenoids

1. Transfer 250 μ l of culture to a 2 ml microcentrifuge tube with screw cap.
2. Collect cells by centrifugation at $11,000 \times g$ for 2 min. Remove the supernatant avoiding pellet disturbance.

3. Add 250 μl of 425–600 μm acid-washed glass beads (Sigma) and 1 ml of dodecane.
4. Lyse cells using an analog Disruptor Genie[®] Cell Disruptor (Scientific Industries) or equivalent. By trial and error, we determined that by treating the samples twice for 6 min, we ensure maximum cell disruption and carotenoids recovery. However, the disruption time is strain dependent. We encourage the researcher to determine the appropriated disruption time for the strain-of-interest.
5. Separate cell debris and glass beads from the supernatant by centrifugation at $18,000 \times g$ for 2 min. If pelleted cells remain pigmented, this indicates that extraction was not complete; therefore, the disruption time should be increased.
6. Transfer 200 μl of the supernatant to a Corning[®] 96-well clear-bottom plate for quantification.
7. Ideally, HPLC analysis should be used to obtain exact quantifications of carotenoids (*see Note 5*). However, in the absence of available tools, a spectrophotometer or microplate reader (TECAN Infinites[®] or similar) can be used to quantify total carotenoids using a survey scan from OD₃₅₀ to OD₅₅₀, in order to determine any shifts in the spectrum. The relative total carotenoids production is determined by calculating the area under the curve of the survey scan, using the parental strain as reference. The maximum absorbance of β -carotene occurs at 454 nm, thus quantification of β -carotene is determined by the absorption at 454 nm. It is vital to have a standard curve for β -carotene quantification. This can be generated using commercially available β -carotene (Enzo Life Sciences) at OD₄₅₄.

4 Notes

1. *It is crucial to determine the level of toxicity of the bio-product of interest.* In this chapter we described how to compensate for the metabolic load imposed by the biosynthesis of carotenoids; carotenoids are generally not toxic to the cells. In this case, we were able to design an effective selective pressure to direct evolution towards the desired phenotype. If the compound-of-interest not only imposes a metabolic burden, but is also toxic to the microorganism, then the described methodology may not be adequate to overcome those issues. In those cases, it is recommended to first increase host tolerance to the toxic bio-product, and once a tolerant strain is obtained, the described methodology can be implemented to increase its production.
2. Chemical mutagenesis can be used to accelerate the evolutionary process by increasing the mutation rate of the system.

However, it is recommended to use it cautiously to avoid getting too many mutations per strain to prevent beneficial mutants with improved productivity to acquire deleterious mutations.

3. It is imperative to periodically track evolving populations for microbial and fungal contamination. Commonly used techniques for this purpose includes: PCR genotyping, Chromocult[®] Coliform Agar, Gram staining, and visual inspection of culture under microscope.
4. Maintenance of sterile conditions is mandatory to prevent contamination. Every accessory in direct contact with the evolving populations must be sterilized.
5. The protocol for carotenoids quantification we described may not be accurate if different carotenoids with overlapping absorbance spectra are produced; although it is a simple method for qualitative assessment of strain performance. For more accurate quantification of carotenoids, HPLC [19, 20] can be used.

References

1. Reyes LH, Gomez JM, Kao KC (2014) Improving carotenoids production in yeast via adaptive laboratory evolution. *Metab Eng* 21:26–33. doi:[10.1016/j.ymben.2013.11.002](https://doi.org/10.1016/j.ymben.2013.11.002)
2. Almario MP, Reyes LH, Kao KC (2013) Evolutionary engineering of *Saccharomyces cerevisiae* for enhanced tolerance to hydrolysates of lignocellulosic biomass. *Biotechnol Bioeng* 110(10):2616–2623. doi:[10.1002/bit.24938](https://doi.org/10.1002/bit.24938)
3. Reyes LH, Almario MP, Winkler J, Orozco MM, Kao KC (2012) Visualizing evolution in real time to determine the molecular mechanisms of n-butanol tolerance in *Escherichia coli*. *Metab Eng* 14(5):579–590. doi:[10.1016/j.ymben.2012.05.002](https://doi.org/10.1016/j.ymben.2012.05.002)
4. Bentley WE, Mirjalili N, Andersen DC, Davis RH, Kompala DS (1990) Plasmid-encoded protein: the principal factor in the “metabolic burden” associated with recombinant bacteria. *Biotechnol Bioeng* 35(7):668–681. doi:[10.1002/bit.260350704](https://doi.org/10.1002/bit.260350704)
5. Rozkov A, Avignone-Rossa CA, Ertl PF, Jones P, O’Kennedy RD, Smith JJ, Dale JW, Bushell ME (2004) Characterization of the metabolic burden on *Escherichia coli* DH1 cells imposed by the presence of a plasmid containing a gene therapy sequence. *Biotechnol Bioeng* 88(7):909–915. doi:[10.1002/bit.20327](https://doi.org/10.1002/bit.20327)
6. Verwaal R, Wang J, Meijnen JP, Visser H, Sandmann G, van den Berg JA, van Ooyen AJ (2007) High-level production of beta-carotene in *Saccharomyces cerevisiae* by successive transformation with carotenogenic genes from *Xanthophyllomyces dendrorhous*. *Appl Environ Microbiol* 73(13):4342–4350. doi:[10.1128/AEM.02759-06](https://doi.org/10.1128/AEM.02759-06)
7. Li Q, Sun Z, Li J, Zhang Y (2013) Enhancing beta-carotene production in *Saccharomyces cerevisiae* by metabolic engineering. *FEMS Microbiol Lett* 345(2):94–101. doi:[10.1111/1574-6968.12187](https://doi.org/10.1111/1574-6968.12187)
8. Xie W, Lv X, Ye L, Zhou P, Yu H (2015) Construction of lycopene-overproducing *Saccharomyces cerevisiae* by combining directed evolution and metabolic engineering. *Metab Eng* 30:69–78. doi:[10.1016/j.ymben.2015.04.009](https://doi.org/10.1016/j.ymben.2015.04.009)
9. Nielsen J (2009) Systems biology of lipid metabolism: from yeast to human. *FEBS Lett* 583(24):3905–3913. doi:[10.1016/j.febslet.2009.10.054](https://doi.org/10.1016/j.febslet.2009.10.054)
10. Xu X, Jin W, Jiang L, Xu Q, Li S, Zhang Z, Huang H (2016) A high-throughput screening method for identifying lycopene-overproducing *E. coli* strain based on an antioxidant capacity assay. *Biochem Eng J* 112:277–284. doi:[10.1016/j.bej.2016.04.032](https://doi.org/10.1016/j.bej.2016.04.032)
11. Wang TY, Tsai YH, Yu IZ, Chang TS (2016) Improving 3’-hydroxygibberic acid production in recombinant *Pichia pastoris* using periodic hydrogen peroxide-shocking strategy. *J Microbiol Biotechnol* 26(3):498–502. doi:[10.4014/jmb.1509.09013](https://doi.org/10.4014/jmb.1509.09013)

12. Kao KC, Sherlock G (2008) Molecular characterization of clonal interference during adaptive evolution in asexual populations of *Saccharomyces cerevisiae*. *Nat Genet* 40(12):1499–1504. doi:[10.1038/ng.280](https://doi.org/10.1038/ng.280)
13. Bahler J, Wu JQ, Longtine MS, Shah NG, McKenzie A 3rd, Steever AB, Wach A, Philippsen P, Pringle JR (1998) Heterologous modules for efficient and versatile PCR-based gene targeting in *Schizosaccharomyces pombe*. *Yeast* 14(10):943–951. doi:[10.1002/\(SICI\)1097-0061\(199807\)14:10<943::AID-YEA292>3.0.CO;2-Y](https://doi.org/10.1002/(SICI)1097-0061(199807)14:10<943::AID-YEA292>3.0.CO;2-Y)
14. Winkler J, Reyes LH, Kao KC (2013) Adaptive laboratory evolution for strain engineering. *Methods Mol Biol* 985:211–222. doi:[10.1007/978-1-62703-299-5_11](https://doi.org/10.1007/978-1-62703-299-5_11)
15. Boeke JD, LaCroutte F, Fink GR (1984) A positive selection for mutants lacking orotidine-5'-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance. *Mol Gen Genet* 197(2):345–346
16. Switala J, Loewen PC (2002) Diversity of properties among catalases. *Arch Biochem Biophys* 401(2):145–154. doi:[10.1016/S0003-9861\(02\)00049-8](https://doi.org/10.1016/S0003-9861(02)00049-8)
17. Martins D, English AM (2014) Catalase activity is stimulated by H₂O₂ in rich culture medium and is required for H₂O₂ resistance and adaptation in yeast. *Redox Biol* 2:308–313. doi:[10.1016/j.redox.2013.12.019](https://doi.org/10.1016/j.redox.2013.12.019)
18. Winkler JD, Garcia C, Olson M, Callaway E, Kao KC (2014) Evolved osmotolerant *Escherichia coli* mutants frequently exhibit defective N-acetylglucosamine catabolism and point mutations in cell shape-regulating protein MreB. *Appl Environ Microbiol* 80(12):3729–3740. doi:[10.1128/AEM.00499-14](https://doi.org/10.1128/AEM.00499-14)
19. Breithaupt DE (2004) Simultaneous HPLC determination of carotenoids used as food coloring additives: applicability of accelerated solvent extraction. *Food Chem* 86(3):449–456. doi:[10.1016/j.foodchem.2003.10.027](https://doi.org/10.1016/j.foodchem.2003.10.027)
20. Kaiser P, Surmann P, Vallentin G, Fuhrmann H (2007) A small-scale method for quantitation of carotenoids in bacteria and yeasts. *J Microbiol Methods* 70(1):142–149. doi:[10.1016/j.mimet.2007.04.004](https://doi.org/10.1016/j.mimet.2007.04.004)

Part IV

Learnings from Design-Build-Test-based Projects

Chapter 21

Two-Scale ^{13}C Metabolic Flux Analysis for Metabolic Engineering

David Ando and Hector Garcia Martin

Abstract

Accelerating the Design–Build–Test–Learn (DBTL) cycle in synthetic biology is critical to achieving rapid and facile bioengineering of organisms for the production of, e.g., biofuels and other chemicals. The Learn phase involves using data obtained from the Test phase to inform the next Design phase. As part of the Learn phase, mathematical models of metabolic fluxes give a mechanistic level of comprehension to cellular metabolism, isolating the principle drivers of metabolic behavior from the peripheral ones, and directing future experimental designs and engineering methodologies. Furthermore, the measurement of intracellular metabolic fluxes is specifically noteworthy as providing a rapid and easy-to-understand picture of how carbon and energy flow throughout the cell. Here, we present a detailed guide to performing metabolic flux analysis in the Learn phase of the DBTL cycle, where we show how one can take the isotope labeling data from a ^{13}C labeling experiment and immediately turn it into a determination of cellular fluxes that points in the direction of genetic engineering strategies that will advance the metabolic engineering process.

For our modeling purposes we use the Joint BioEnergy Institute (JBEI) Quantitative Metabolic Modeling (jQMM) library, which provides an open-source, python-based framework for modeling internal metabolic fluxes and making actionable predictions on how to modify cellular metabolism for specific bioengineering goals. It presents a complete toolbox for performing different types of flux analysis such as Flux Balance Analysis, ^{13}C Metabolic Flux Analysis, and it introduces the capability to use ^{13}C labeling experimental data to constrain comprehensive genome-scale models through a technique called two-scale ^{13}C Metabolic Flux Analysis (2S- ^{13}C MFA) [1]. In addition to several other capabilities, the jQMM is also able to predict the effects of knockouts using the MoMA and ROOM methodologies. The use of the jQMM library is illustrated through a step-by-step demonstration, which is also contained in a digital Jupyter Notebook format that enhances reproducibility and provides the capability to be adopted to the user's specific needs. As an open-source software project, users can modify and extend the code base and make improvements at will, providing a base for future modeling efforts.

Key words Flux analysis, ^{13}C Metabolic flux analysis, Omics data, Predictive biology

Electronic supplementary material: The online version of this chapter (https://doi.org/10.1007/978-1-4939-7295-1_21) contains supplementary material, which is available to authorized users.

Michael Krogh Jensen and Jay D. Keasling (eds.), *Synthetic Metabolic Pathways: Methods and Protocols*, Methods in Molecular Biology, vol. 1671, https://doi.org/10.1007/978-1-4939-7295-1_21, © Springer Science+Business Media, LLC 2018

1 Introduction

The capability to change an organism's DNA through genetic engineering has radically changed the nature of biology in the last few decades. Synthetic biology was born in the twenty-first century as a reinterpretation of genetic engineering applying systematic design [2] and traditional engineering principles. One of those engineering principles is the Design–Build–Test–Learn (DBTL) cycle: a loop used recursively to obtain a design that satisfies the desired specifications [3]. The DBTL cycle starts with the design (D) of the biological system to produce the desired outcome. That design is built (B) in the next phase from DNA parts and an appropriate microbial chassis using synthetic biology tools. The next phase involves testing (T) whether the biological system performs as desired in the original design using a variety of assays (e.g., production measurement or/and omics profiling). It is extremely unlikely that the first design behaves as desired, and further attempts will most likely be needed to meet the desired specification. It would be desirable not to do these posterior attempts randomly, but rather to use the data generated in previous rounds to converge towards engineering goals more quickly. This phase is called the learn (L) phase of the DBTL cycle and is, arguably, the hardest and most weakly supported step in current metabolic engineering practice [3]. However, given the ever increasing amounts of data provided by the postgenomics revolution and current increasingly available high-throughput workflows, there is an imperative need to efficiently use the test data to provide *actionable* items for metabolic engineers: i.e. suggestions that can be acted upon with available tools and protocols (e.g., to change a particular gene's RBS or knock out a particular gene in order to increase a specific flux or to accelerate growth).

In this chapter we will show how to use metabolomic data obtained from ^{13}C labeling experiments to generate actionable items to increase acetate production in *E. coli*. We will use the JBEI Quantitative Metabolic Modeling library (jQMM) [4] to calculate cellular fluxes and make predictions. The jQMM library is currently capable of measuring and predicting internal metabolic fluxes using three different techniques: ^{13}C Metabolic Flux Analysis (^{13}C MFA) [5], Flux Balance Analysis (FBA) [6], and two-scale ^{13}C Metabolic Flux Analysis (2S- ^{13}C MFA) [1]. First we will provide a brief description of ^{13}C labeling experiments, which provide the needed experimental data, and which consist of cellular cultures in which the feed (e.g., glucose) is labeled with carbon atoms that have an extra neutron (i.e. carbon isotopes) at selected positions. We will then succinctly describe how to measure the ensuing labeling in the metabolites in the studied cells (these are steps more appropriately described as part of the test phase). Next we describe

in detail how to use the labeling data from different metabolites in the cell to infer what the cell's internal metabolic fluxes are through a technique called two-scale ^{13}C Metabolic Flux Analysis (2S- ^{13}C MFA). 2S- ^{13}C MFA introduces the capability to use experimental ^{13}C labeling data to constrain comprehensive *genome-scale* models (rather than small models of central metabolism as done with traditional ^{13}C MFA) by taking into account the system-wide balances of metabolites [7]. Finally, we will show how to use the COBRA (Constraint-Based Reconstruction and Analysis) [8] methods of MoMA (Minimization of Metabolic Adjustment) [9] and ROOM (Regulatory On/Off Minimization) [10] to predict, based on the measured flux profiles, which gene knockouts will increase acetate production in *E. coli*.

2 Materials

The general workflow for determining and plotting of metabolic fluxes is shown in Fig. 1. First, a microbial culture is grown with ^{13}C -labeled glucose. Next, mass spectroscopy is used for the analysis of the distribution of ^{13}C in metabolites taken from cell culture

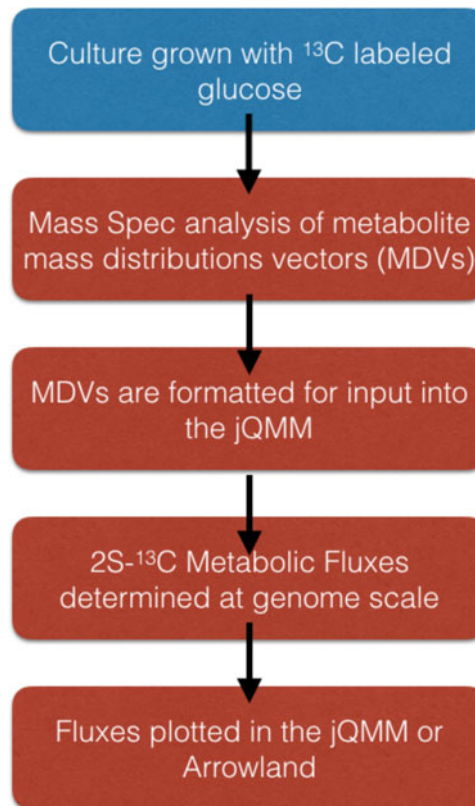


Fig. 1 Overview of the workflow for ^{13}C Two-Scale Metabolic Flux Analysis. The Test phase of the workflow is in *blue* while the Learn phase steps are in *red*

to create Mass Distributions Vectors (MDVs) [11] for each measured metabolite, which give the relative frequency of isotopomers. Each isotopomer of a metabolite has a different number of carbons with an extra neutron, and each isotopomer can have the heavy carbon(s) at any carbon(s) located within the metabolite. The different labeling patterns for the measured metabolites will ultimately provide enough information to determine the internal flux profile [1]. MDVs then need to be formatted for input into the jQMM, so that metabolic fluxes can be calculated at the genome scale using the $2S\text{-}^{13}\text{C}$ methodology. For genome-scale models, there are thousands of reactions, so fluxes can be more easily understood via plotting in Arrowland software (<http://public-arrowland.jbei.org>) or within the jQMM.

2.1 Computational Requirements

Procurement of a modern desktop computer or server system that is capable of running heavy computational loads for extended periods of time is necessary. The more available cores/CPU's in the system the better, as the jQMM library is parallelized and can leverage additional cores/CPU's to reduce computation time. Windows, Mac, and Linux operating systems are all compatible with running python code, which the jQMM is written in. Error-correcting code RAM (ECC RAM) is recommended to reduce the probability of faulty calculations over lengthy computations due to errors in writing and reading to RAM memory. Xeon processors from Intel for example, are specially designed for long continuous computations at high CPU load, unlike the consumer series of desktop processors sold by Intel such as the i7 and i5 series (circa 2016).

$2S\text{-}^{13}\text{C}$ MFA computations require the following:

1. At least 32 GB of RAM and 500 GB of free disk space.
2. Ownership of both GAMS and CONOPT solver licenses.
3. Python version 2.7 installed.

2.2 Software Installation and Configuration

Installation of the required libraries for using the jQMM can be done in two different ways: the traditional way, which includes a self-installation on the hardware at hand, or via the use of a pre-configured and preinstalled Docker container (<http://www.docker.com>). We recommend using our pre-configured Docker container for use of the jQMM as the flux modeling environment will be immediately usable. However, for expert users or those who wish to use a custom installation, a self-installation is readily achieved.

2.2.1 Self-Installation

First install the following jQMM dependencies to the Python 2.7 environment:

1. libSBML: available at <http://sbml.org/Software/libSBML>
2. matplotlib: available at <http://matplotlib.org/users/installing.html>

3. numpy: available at <http://www.scipy.org/scipylib/download.html>
4. jupyter: available at <http://jupyter.org/>

Download the jQMM library from <https://github.com/JBEI/jqmm> and unpack the downloaded file. The jQMM is best used through an interactive python (iPython) notebook server called Jupyter Notebook (<http://jupyter.org/>), which is a web application that allows for the interactive execution, visualization, and documentation of python code. This integrated computing format greatly enhances reproducibility of results. Next, the jQMM library can be used by logging into the local Jupyter server using a web browser and navigating to the jQMM folder, and then running some of the example Jupyter notebooks contained within the jQMM. If desired, the Jupyter server can be run directly from the command line within a linux terminal via the command “jupyter notebook”. GAMS and CONOPT licenses are needed and must be obtained separately.

2.2.2 Pre-configured and Preinstalled jQMM Library

Docker (<https://docs.docker.com/>) is a technology that is based on Linux containers that allows for building, running, testing, and deploying applications such as the jQMM library into a complete file system that contains everything it needs to run: code, runtime, system tools, system libraries, and supporting data files. This guarantees that it will always run correctly and in the same way, regardless of the system environment it is running in. The jQMM docker container can be run on virtually any cloud computing service such as AWS (Amazon Web Services), Google Cloud Platform, and Microsoft Azure. Additionally, the jQMM Docker container can be conveniently run on a personal computer running either Microsoft Windows or the Mac operating system, although we discourage this practice for anything other than training purposes given how slowly the jQMM will run on personal computers.

If choosing to run the jQMM docker container on a web-based platform one avoids the need of having to purchase an expensive high-performance server system. Pricing for cloud computing services is typically based on usage, which allows for the ability to automatically adjust the usage of computational services to as much or as little as needed, at any time. When choosing a type of instance to use on a cloud based system we recommend instances which focus on computational speed and not RAM size or disk drive access speed. On the AWS, this includes the instances of type M4 and C4, with the C4 instance currently featuring the highest performing processors and the lowest price/compute performance ratio offered by AWS (circa 2016).

The jQMM docker container (available for download at <https://github.com/JBEI/jqmm>) has all of the software needed to run the jQMM library preinstalled and pre-configured to work

out of the box. The GAMS and CONOPT solvers (<http://www.gams.com/>, <http://www.conopt.com/>), which are preinstalled, are required to do flux analysis, but their usage requires purchase of a GAMS license and CONOPT license separately. Once these licenses are included with the GAMS and CONOPT solver installations the jQMM docker container will be fully functional.

2.3 ¹³C Labeling Experiments and Mass Spectrometry Data Analysis

The initial step for any ¹³C-based metabolic flux experiment is performing a ¹³C labeling experiment with the organism of interest. Since this step is more related to the Test phase than the Learn phase of the DBTL cycle, we will only give a brief description, and refer the reader to previous protocols [12, 13]. Cultures must use minimal media, and can be grown using different types of labeling for the feed (*see Note 1*). A common choice is to use 20% normal glucose and 80% 1-¹³C glucose, for example. Briefly, we recommend that culture samples are prepared by taking an aliquot of the cultured cells which were grown with ¹³C glucose and filtering with a 0.45- μ m pore-sized filter. The filter is then washed with Milli-Q water to remove the cultured cells which are then placed in methanol at 4 °C to halt metabolism. A solution at a ratio of 4 mL of chloroform to 1.6 mL of Milli-Q water is mixed with the filtered cells and then centrifuged at $2300 \times g$ for 5 min at 4 °C. To remove high-molecular weight compounds the methanol layer is extracted and then passed through a Millipore 5-kDa cutoff filter via centrifugation. Finally, the filtrate is lyophilized and then dissolved in Milli-Q water before analysis on a mass spectrometry instrument.

The quantification of relative cellular metabolite isotopomer concentration, consisting of MDVs for metabolites in cellular metabolism (*see Fig. 2* for an example), is done by determining the

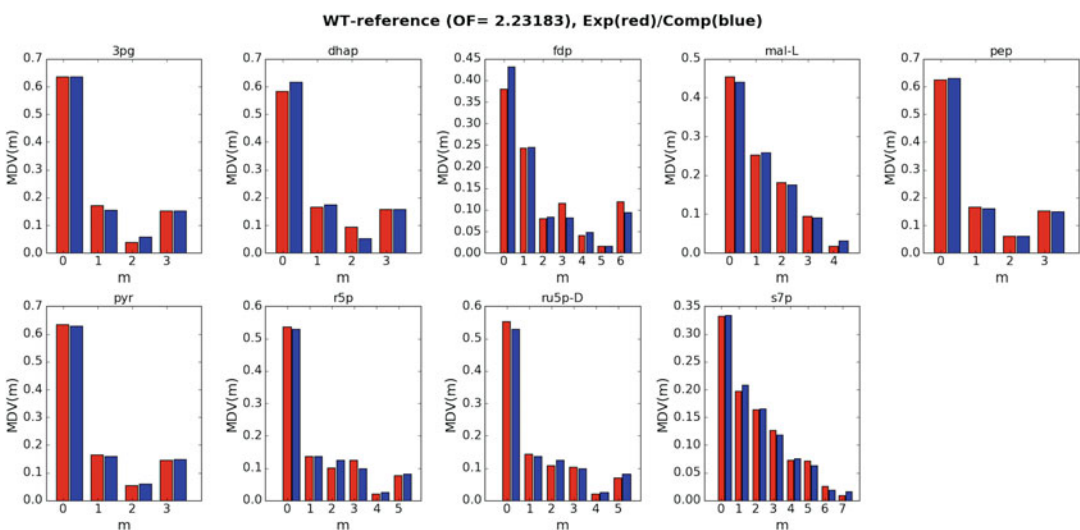


Fig. 2 Plot of wild-type *E. coli* MDVs from Toya et al. [12]: Experimentally measured MDVs are plotted in red bars for each metabolite, while the MDVs implied by the predicted fluxes are shown in the blue bars. Computational and experimental data match closely implying that the model is quantitatively correct

relative concentration of each metabolite's isotopomers via mass spectrometry. Once chromatograms from a mass spectrometry instrument have been analyzed and integrated such that the relative frequency of each metabolite's mass distribution has been determined, these data need to be prepared and formatted for input into the jQMM library so that metabolic fluxes can be determined.

2.4 Input Data Preparation

2.4.1 ¹³C Labeling Data

Once the measured MDVs have been determined and normalized such that their sum is equal to one for each metabolite, they have to be entered into a text document with the following format:

```
Amino acid Mass distribution
      m0    m1    m2    m3    m4    m5  m6  m7  m8
3pg   M-0  0.387  0.408  0    0.204  -  -  -  -
ala-L M-0  0.382  0.379  0.059  0.178  -  -  -  -
asp-L M-0  0.297  0.429  0.273  0    0    -  -  -
dhap  M-0  0.348  0.464  0    0.186  -  -  -  -
```

The first two lines should remain fixed and are ignored by the jQMM library, while the following lines need to include the MDV information for every metabolite for which isotopomer data exists in the following format:

1. First, the metabolite name using the metabolite abbreviation used in the BIGG database (<http://bigg.ucsd.edu>) is specified and followed by a tab character.
2. Next, the tab-separated relative frequencies of each metabolite isotopomer are specified. The sum of all isotopomer frequencies needs to add to 1 for each individual metabolite.
3. Isotopomers, which do not exist for a particular metabolite, should be represented by a “-” (dash), and must not be entered in as a zero value, which would indicate that such an isotopomer does exist but is not present in the sample.

2.4.2 Choose, Edit, and Define the Metabolic Model

To define a metabolic model for the jQMM to use in the modeling of the particular organism being studied, navigate to the BIGG database (<http://bigg.ucsd.edu>) and download a SBML version of a metabolic model which is appropriate to the problem being studied. Smaller models run much faster in the jQMM, while more comprehensive genome-scale models are necessary for problems involving peripheral metabolism or which include the 2S-¹³C MFA analysis methodology. Retooling of the jQMM library code may be required for the library to understand metabolite names which do not follow the naming convention used in the iJR904 metabolic model [14] format, with sample code already included for using metabolite names from the iJO1366 [15] and iAF1260 [16] models, which is located in the “sbmlio.py” python code file located in the jQMM “code/core” directory. Network

reactions, which come from a heterologous pathway that has been engineered into a microbe, should be manually entered into the downloaded SBML model.

2.4.3 Defining Exchange Reaction Fluxes

Both measured fluxes through exchange reactions and the biomass flux that correspond to the time that a sample was taken for ^{13}C isotopomer analysis are detailed in a ‘FLUX.txt’ file. A sample exchange flux file is as follows:

```
BiomassEcoli: 0.70 [==] 0.76
GLCpts:      11.1 [==] 11.1
EX_glc(e):  -11.1 [==] -11.1
EX_ac(e):    2.6 [==] 2.6
```

The Biomass flux is in units of l/h. and is normalized to equal the growth rate while all other exchange fluxes are in units of mMol/gdw/h (millimoles/grams of dry weight/hour). The glucose uptake rate can be measured by the HPLC determination of glucose concentration at two different times around the ^{13}C analysis sample time. If the glucose concentration is measured as g_1 and g_2 (in millimoles/volume) at times t_1 and t_2 , the glucose flux can be approximated as $(g_2 - g_1)/\text{gdw}_1/(t_2 - t_1)$, where gdw_1 is the grams of dry weight of cells per unit volume at time t_1 . Similarly, HPLC measurements can be used to determine fluxes of acetate, lactate, and other organic acids excreted by the cell.

2.4.4 Defining Carbon Transitions in the Metabolic Model

Atom transitions can be used to represent the fate of each carbon in a reaction [5]. In the example reaction:



The uppercase letters represent the metabolites present in a reaction, while the lowercase letters in parentheses represent the atom transitions. (Note: It is not necessary to follow this convention of using uppercase for metabolites and lowercase for atom transitions. The parentheses delimit the start and end of each atom transition. Any alphabetic, numeric, or underscore character comprising the regular expression [a-zA-Z0-9] can be included in the metabolite names and atom transitions.) Irreversible reactions are denoted by a ‘ \rightarrow ’ or ‘ \Rightarrow ’ arrow and reversible reactions are denoted by a ‘ \leftrightarrow ’ or ‘ \rightleftharpoons ’ arrow. In the jQMM, multiple reactions are separated by carriage returns or by placing a semicolon at the end of each reaction equation.

For example, in:



akg gets split into succoa and co2 in reaction AKGDH, with the first four carbons going to succoa and the remaining carbon going to co2.

For 2S-¹³C-based metabolic flux analysis in the jQMM, one needs to create a ‘REACTIONS.txt’ file which contains the carbon transition information for core reactions in the metabolic network and which has the following format:

```
# Metabolic Exchange fluxes
&MLIST etoh[e] 0
&MLIST ac[e] 0
&MLIST lac-L[e] 0

# Intracellular metabolite fluxes
&MLIST fdp 0
&MLIST dhap 0
&MLIST pep 0
&MLIST r5p 0
&MLIST s7p 0
&MLIST mal-L 0

# Carbon source
&SOURCE glc-D[e]

# Input Reactions
EX_glc(e) glc-D[e] <==> glcDEx abcdef : abcdef
GLCt2 glc-D[e] --> glc-D abcdef : abcdef
HEX1 glc-D --> g6p abcdef : abcdef

# Carbon Transitions
GLCpts glc-D[e] + pep --> g6p + pyr abcdef + ABC : abcdef + ABC
PGI g6p <==> f6p abcdef : abcdef
PFK f6p --> fdp abcdef : abcdef
FBA fdp <==> g3p + dhap CBAabc : abc + ABC
```

Reactions which must be specified in the ‘REACTIONS.txt’ file, in terms of carbon atom transition information, are those that utilize metabolites for which ¹³C isotopomer data are input into the jQMM library and for reactions which are considered to be at the core of the metabolic network. Finally, this file also specifies a metabolite which serves as the ¹³C carbon source (typically glucose, i.e. glc-D[e], via the &SOURCE command) and input reactions which bring this ¹³C-labeled metabolite into the cell.

2.4.5 Defining Feed Labeling

The type of labeled glucose used in the experiment, together with its concentration relative to the amount of unlabeled glucose, is detailed in a ‘FEED.txt’ file. A sample feed specification file contains the following line:

```
0.4% Glucose: 30% 1-C 20% U 50% UN
```

The 0.4% at the beginning of a feed definition specifies the total glucose percentage of the initial cell culture. The percentage of I-C glucose, which has its first carbon atom labeled, is specified next, together with the percentage of U glucose, which is uniformly labeled among all the glucose carbon atoms, and finally the percentage of normal glucose which is completely unlabeled (UN).

3 Methods

The example Jupyter notebook which is included as an attachment in the online version of this protocol, and which is also reproduced in this section, contains a description of how to use the FluxModels module in the jQMM to do 2S-¹³C MFA and then predict the outcomes on acetate production of different reaction knockouts. The notebook provides a convenient way to reproduce results and is easily modified to fit the user's specific needs (*see Note 2*). It is broken into six different steps for turning experimental data into actionable predictions for increasing a targeted biochemical via genetic engineering:

1. Gathering input data
2. Creating the Reaction Network
3. Creating the two-scale metabolic model
4. Calculating internal metabolic fluxes through 2S-¹³C MFA
5. Visualizing flux profiles
6. Predicting which genes to knock out using MoMA and ROOM

-- Jupyter Notebook Start --

Predicting KO outcomes with 2S-¹³C MFA and COBRA methods

This Jupyter notebook presents a computable step-by-step description of how to use metabolite data from ¹³C labeling experiments to produce actionable insights to improve acetate production in *E. coli*.

0. Setup

The first step involves specifying the correct path for the library:

```
In[1]:
%matplotlib inline
import sys, os
path = "/scratch/user"
pythonPath = path + "/quantmodel/code/core"
if pythonPath not in sys.path:
    sys.path.append(path + '/quantmodel/code/core')
os.environ["QUANTMODEL_PATH"] = path + '/quantmodel'
```


We then need to import the needed classes for the notebook:

```
In[2]:
from IPython.display import SVG
import FluxModels as FM
import enhancedLists, ReactionNetworks, predictions, copy,
core
```

and then move to a defined working directory where output and intermediate files will be kept:

```
In[3]:
cd /scratch/user.working_dir/tests

Out[3]:
/scratch/user.working_dir/tests
```

1. Gathering input data

As part of the test (T) phase of the DBTL cycle, we gather all the relevant experimental data from the ^{13}C labeling experiments (*see* Subheading 2). These data involve:

1. A base *genome-scale model* that will act as the reference for all other data types [14].
2. *Exchange fluxes* containing the measured fluxes of metabolites being exchange by cells with the environment.
3. *Transition information* on the fate of each carbon in the core reaction network [17].
4. *Metabolite labeling* information in the form of Mass Distribution Vectors (MDVs).
5. *Metabolite labeling error* information.
6. *Feed labeling* information on the type of labeled glucose the cell culture was fed.

Discussion of these data types can be seen in Subheading 2.

For this demonstration, we will use the data from Toya et al. [12]:

```
In[4]:
datadir = os.environ['QUANTMODELSPATH'] + '/data/tests/Toya2010/
2S/wt5h/'
strain = 'wt5h'
BASEfilename = datadir + 'EciJR904TKs.xml'
FLUXESfilename = datadir + 'FLUX'+strain+'.txt'
TRANSITIONSfilename = datadir + 'REACTIONS'+strain+'.txt'
MSfilename = datadir + 'GCMS'+strain+'.txt'
MSSTDfilename = datadir + 'GCMSerr'+strain+'.txt'
FEEDfilename = datadir + 'FEED'+strain+'.txt'
```

2. Creating the Reaction Network

Once we have gathered all the needed input files, we can condense all this information into a single sbml file. We will do this using a reaction network from the ReactionNetworks module in the jQMM library. A reaction network contains all information related to the metabolic reaction network used for the simulation:

```
In[5]:
# Load initial SBML file
reacNet = ReactionNetworks.TSReactionNetwork(BASEfilename)
# Add Measured fluxes
reacNet.loadFluxBounds(FLUXESfilename)
# Add carbon transitions
reacNet.addTransitions(TRANSITIONSfilename,translate2SBML=True)
# Add measured labeling information
reacNet.addLabeling(MSfilename,'LCMS',MSSTDfilename,min
STD=0.001)
# Add feed labeling information
reacNet.addFeed(FEEDfilename)
# Limit fluxes to 500
reacNet.capFluxBounds(500)
# Create sbml file to store the two-scale model.
# All input files are combined in a tuple of the type:
# (fileName, string of contents)
SBMLfile = ('EciJR904TKs'+strain+'TS.xml',reacNet.write('to
String'))
```

3. Creating the two-scale metabolic model

The next step is to use the SBML file we just created to create a two-scale model [1] that we will use to calculate fluxes through 2S-¹³C MFA:

```
In[6]:
TSmodel = FM.TwoSC13Model(('EciJR904TKs'+strain+'TS.xml',
reacNet.write('toString')))
```

TSmodel now contains all the information needed to calculate fluxes along with the methods to do this calculation and other analysis [1].

4. Calculating internal metabolic fluxes through 2S-¹³C MFA

We can now use the *findFluxesRanges* method in *TSmodel* to find the fluxes that best fit the experimentally obtained metabolite labeling data (MDVs) and find the ranges of fluxes compatible with this labeling data and the corresponding experimental error:

```
In[7]:
fluxNames = TSmodel.reactionNetwork.C13ReacNet.reactionList.
getReactionNameList(level=1)
TSresult = TSmodel.findFluxesRanges(Nrep=30, fluxNames=flux
Names, procString='proc')
```

Nrep represents the number of replicates used for the calculation. Since the problem to be solved is a nonconvex problem there is no guarantee that a single run will find the best global fit. Hence we run 30 independent processes and keep the one that best fits the data. *fluxNames* indicates the fluxes for which full flux confidence intervals will be calculated. *procString* indicates that the data (for this case) needs no derivatization correction.

We can check how accurate the model is by comparing the measured labeling distribution (MDVs, red) with the one predicted through the computational model (blue) by using the *plotExpvsCompLabelFragment* method:

```
In[8]:
%%time
TSresult.plotExpvsCompLabelFragment(titleFig='WT-reference')
```

```
Out[8]:
CPU times: user 824 ms, sys: 12 ms, total: 836 ms
Wall time: 831 ms
```

See Fig. 2.

or by using *plotExpvsCompLabelXvsY*, if we prefer to see these fits as an *X* vs. *Y* plot:

```
In[9]:
TSresult.plotExpvsCompLabelXvsY(titleFig = 'WT-reference')
```

```
Out[9]:
```

See Fig. 3.

The closeness of the fit data and the experimental data validate the use of this model.

Results are stored in a reaction network inside *TSresult* and can be explored through the *reactionList* methods.

For example, we can print the desired fluxes:

```
In[10]:
TSresult.reactionNetwork.reactionList.printFluxes(brief
="True", names="exchange")
```

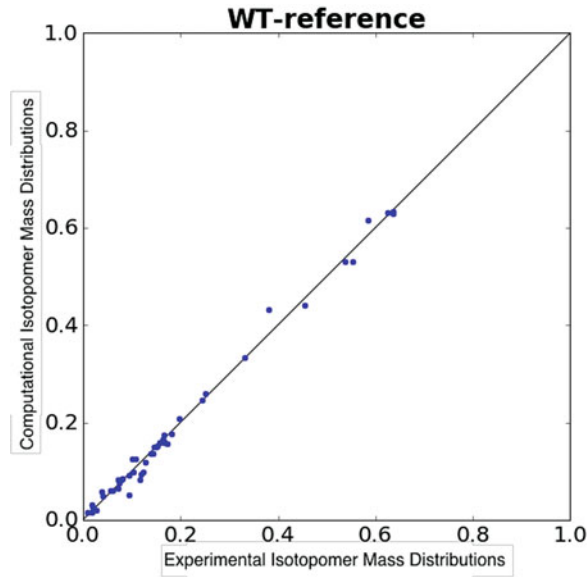


Fig. 3 Plot of experimentally measured MDVs versus the computationally predicted MDVs in an x - y plot (different way to plot data in Fig. 2), which demonstrates that experimental and predicted MDVs are comparable

```
Out[10]:
EX_h2o_e_: 43.8140702078
EX_co2_e_: 26.0397076138
EX_o2_e_: -24.3853340558
EX_h_e_: 12.3121388881
EX_glc_e_: -11.7
EX_nh4_e_: -8.93782429
EX_ac_e_: 4.3
BiomassEcoli: 0.83
EX_pi_e_: -0.75665871
EX_acald_e_: 0.22513003
EX_so4_e_: -0.19343897
EX_succ_e_: 0.145316144053
EX_glyc1t_e_: 0.0415
EX_urea_e_: 0.03486
```

Or we can retrieve the computationally predicted labeling distribution or the experimentally measured one:

```
In[11]:
TSresult.EMUlabel['pep']

Out[11]:
array([ 0.63083, 0.15962, 0.06024, 0.14931])

In[12]:
TSresult.fragDict['pep'].mdv

Out[12]:
array([ 0.624, 0.165, 0.06 , 0.151])
```

An important test to make sure that the assumptions used in the $2\text{S-}^{13}\text{C}$ MFA properly hold is the External Labeling Variability Analysis (ELVA) [1] test. This test checks that the reactions for which no carbon transition information was available do not significantly distort the flux solution obtained.

```
In[13]:
resultELVA = TSmodel.ELVA(TSresult)
```

ELVA results can be plotted in an x - y graph showing the experimentally determined isotope labeling which defines a confidence interval that represents the maximum possible difference in labeling that could be attributed to non-core reactions for the current solution. The reactions that contribute an unacceptable amount of uncertainty are then added to the core set and the procedure can be repeated as necessary, until a core set of reactions is found which fully justifies the two-scale approximation. In this example, all reactions have only small fluctuations in predicted computational labeling.

```
In[14]:
resultELVA.plotExpvsCompLabelxvsy(titleFig="WT",outputFileName="ELVAComparisonWT.txt",save="ELVA-W.eps")
Out[14]:
```

See Fig. 4.

The error bars in the y axis (computational error) are of the same order of magnitude as the experimental error, hence justifying the two-scale assumption [1].

5. Visualizing flux profiles

Once the metabolic fluxes have been calculated they can be understood visually via their plotting on a flux map. In the jQMM library fluxes can be plotted via the commands:

```
In[15]:
TSresult.drawFluxes('wt.svg',svgInFileName='TOYAexp.svg',
norm='EX_glc_e_')
Out[15]:
svgin:
/scratch/david.ando/quantmodel/code/core/TOYAexp.svg
```

where 'TOYAexp.svg' is the base flux map contained in the jQMM library [4]. The *drawFluxes()* method will indicate the flux magnitude on the base flux map in two ways: visually by changing the flux arrow width according to the flux magnitude through a reaction, and also numerically by showing the net flux value (with confidence intervals) next to the reaction:

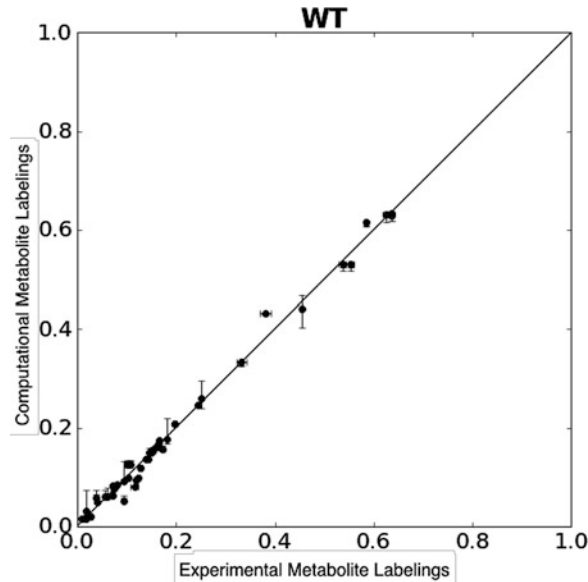


Fig. 4 ELVA plot which shows an x - y graph of the experimentally determined isotope labeling versus the computationally predicted labeling. The *vertical error bars* define the computational error that represent the maximum possible difference in labeling that could be attributed to non-core reactions for the current solution. In this example, all reactions have only small fluctuations in predicted computational labeling, confirming that non-core reactions do not significantly contribute to core metabolite labeling [1]

The command ‘SVG’ displays the flux map in the Jupyter notebook which is contained in the svg file which was saved locally.

```
In[16]:
SVG(filename='wt.svg')

Out[16]:
```

See Fig. 5.

In the near future, one will also be able to display fluxes using the web browser-based flux plotting library Arrowland (<http://public-arrowland.jbei.org>).

6. Predicting which genes to knock out using MoMA and ROOM

So far we have used targeted metabolomic data from ^{13}C labeling experiments to infer the underlying internal metabolic fluxes in the cell. We will now use these inferred fluxes along with two Constraint-Based Reconstruction and Analysis (COBRA) methods to predict which genes to knock out in order to increase the production of acetate. These methods are MoMA (Minimization of Metabolic Adjustment) and ROOM (Regulatory On/Off Minimization). MoMA provides an approximate solution for a suboptimal growth flux state after a knockout has been made to an organism, which is nearest in flux distribution to the unperturbed state [9]. On the other hand, ROOM aims to minimize the number

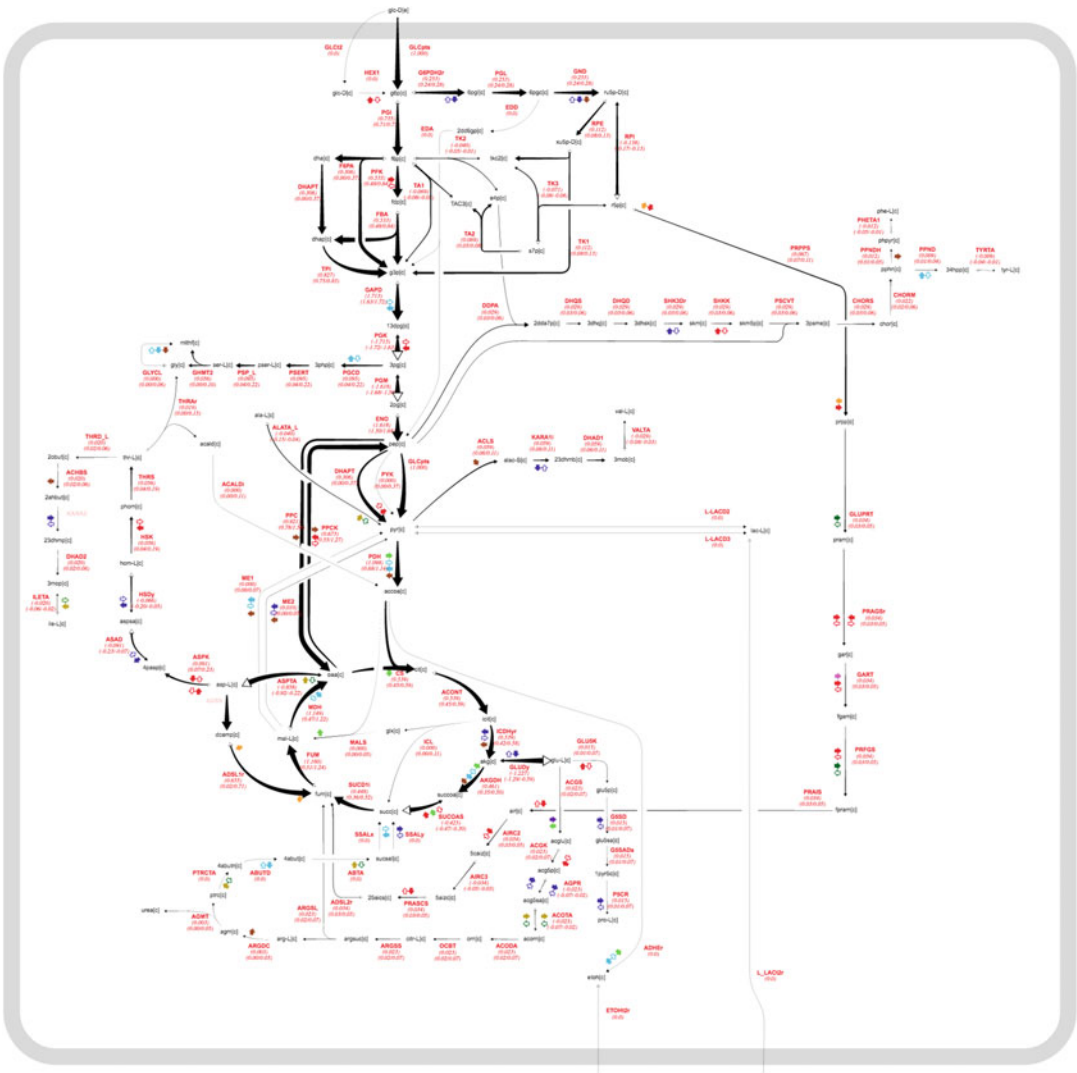


Fig. 5 Plot of 2S-¹³C metabolic fluxes in the jQMM library using the TOYAexp.svg base flux map for a wild-type strain of *E. coli* from Toya et al. [12]. This map can be interactively studied at Arrowland (<https://public-arrowland.jbei.org/>, wt5h)

of significant flux changes with respect to the wild type to predict resultant fluxes from a knockout of a reaction [10].

First we need to specify flexible flux bounds for the final solution in order to avoid biasing the knockout predictions:

```
In[17]:
reactionNetwork = copy.deepcopy(TSmodel.reactionNetwork)
reactionNetwork.changeFluxBounds('GLCpts', core.fluxBounds(0, 25, False)[1])
reactionNetwork.changeFluxBounds('EX_glc_e_', core.fluxBounds
```

```
(-15,-0 ,True,True)[1])
reactionNetwork.changeFluxBounds('BiomassEcoli' ,core.flux
Bounds( 0, 25 ,False)[1])
reactionNetwork.changeFluxBounds('EX_ac_e_' ,core.fluxBounds
(0, 25 ,True,True)[1])
```

Then we can calculate the base flux profiles for MoMA and ROOM:

```
In[18]:
TSresult = TSmodel.findFluxesStdS(Nrep=30,Nrand=10)
```

We then specify a list of reactions to knock out and determine resultant fluxes:

```
In[19]:
KOs = ['RPE','RPI']
```

For reference, we determine the amount of acetate production in the base WT strain:

```
In[20]:
fluxDict = TSresult.reactionNetwork.reactionList.getReaction
Dictionary()
print 'predicted acetate flux = ',fluxDict['EX_ac_e_'].flux.
net.best
Out[20]:
predicted acetate flux = 4.3
```

Perform MOMA and ROOM predictions over the set of specified knockouts:

```
In[21]:
for KO in KOs:
    print KO,'knockout:'
    TS13CMOMA = predictions.predict(TSresult, KO, 'MOMA', reac
tionNetwork.getSBMLString())
    TS13CROOM = predictions.predict(TSresult, KO, 'ROOM', reac
tionNetwork.getSBMLString())
    fluxDict = TS13CMOMA.reactionNetwork.reactionList.getReac
tionDictionary()
    print ' MoMA predicted acetate flux = ',fluxDict['EX_ac_e_'].
flux.net.best
    fluxDict = TS13CROOM.reactionNetwork.reactionList.getReac
tionDictionary()
    print ' ROOM predicted acetate flux = ',fluxDict['EX_ac_e_'].
flux.net.best
    print '—————'
    print ''
```



```
Out[21]:  
RPE knockout:  
MoMA predicted acetate flux = 4.56474059324  
ROOM predicted acetate flux = 6.87162731209  


---

  
RPI knockout:  
MoMA predicted acetate flux = 3.9427989398  
ROOM predicted acetate flux = 4.17  


---


```

As can be observed, knocking out the gene corresponding to the RPE reaction is predicted to increase acetate production by 6.2% according to the MoMA methodology and by 60.0% when using the ROOM methodology. As can be seen with an RPI knock-out, both MoMA and ROOM predict a decline in acetate production.

```
-- Jupyter Notebook End --
```

4 Notes

1. Proper quality control of MDV data is crucial to proper determination of fluxes. Experiments should be designed to include internal controls, and should include several biological and technical replicates.
2. Free open-source software tools, such as the jQMM, provide for universal accessibility and unlimited modification and customization. Overall, we wish that the community can support the jQMM's further development by submitting bug fixes to the github repo (<https://github.com/JBEI/jqmm>) and including any functional extensions that different research groups have achieved.

Acknowledgments

This work was part of the DOE Joint BioEnergy Institute (<http://www.jbei.org>) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

1. H. G. Martín, V. S. Kumar, D. Weaver, and A. Ghosh, A method to constrain genome-scale models with ^{13}C labeling data, 2015, 11(9): e1004363.
2. Kitney R, Freemont P (2012) Synthetic biology - the state of play. *FEBS Lett* 586 (15):2029–2036
3. Nielsen J, Keasling JD (2016) Engineering cellular metabolism. *Cell* 164(6):1185–1197
4. Birkel G, Ghosh A, Vinay K, Weaver D, Ando D, Arkin A, Keasling JD, Martin HG (2017) The JBEI quantitative metabolic modeling library (jQMM): a Python library for modeling microbial metabolism. *BMC Bioinformatics* 18 (1):205
5. Wiechert W (2001) ^{13}C metabolic flux analysis. *Metab Eng* 3(3):195–206
6. Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol* 10 (4):291–305
7. Ghosh A, Ando D, Gin J, Runguphan W, Denby C, Wang G, Baidoo E, Shymansky C, Keasling J, Garcia Martin H (2016) ^{13}C Metabolic flux analysis for systematic metabolic engineering of *S. cerevisiae* for overproduction of fatty acids. *Front Bioeng Biotechnol* 4:76
8. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, Kang J, Hyduke DR, Palsson BØ (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc* 6(9):1290–1307
9. Segrè D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A* 99 (23):15112–15117
10. Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A* 102(21):7695–7700
11. Suthers PF, Burgard AP, Dasika MS, Nowroozi F, Van Dien S, Keasling JD, Maranas CD (2007) Metabolic flux elucidation for large-scale models using ^{13}C labeled isotopes. *Metab Eng* 9(5–6):387–405
12. Toya Y, Ishii N, Hirasawa T, Naba M, Hirai K, Sugawara K, Igarashi S, Shimizu K, Tomita M, Soga T (2007) Direct measurement of isotope-pomer of intracellular metabolites using capillary electrophoresis time-of-flight mass spectrometry for efficient metabolic flux analysis. *J Chromatogr A* 1159(1–2):134–141
13. Zamboni N, Fendt S-M, Rühl M, Sauer U (2009) ^{13}C -Based metabolic flux analysis. *Nat Protoc* 4(6):878–892
14. Reed JL, Vo TD, Schilling CH, Palsson BO (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol* 4(9):R54
15. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BØ (2011) A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism–2011. *Mol Syst Biol* 7(535):535
16. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 3(121):121
17. Antoniewicz MR, Kelleher JK, Stephanopoulos G (2007) Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metab Eng* 9(1):68–86

INDEX

A

Acetogens 151, 157, 159
Actinomycete 165, 167, 177, 181
Adaptive laboratory evolution 149–160, 322,
325, 329
Antibiotic biosynthetic gene clusters 203, 208
Antibiotic resistance 121, 136, 142, 260,
261, 263, 264, 266–268, 270
Antioxidant 320–324
Autotroph 150, 160

B

Bacteria 27–36, 41–43, 46, 57, 138, 157,
159, 163–165, 175, 199, 203, 205, 235, 259,
260, 262, 272, 281, 320, 321
BioDesign automation 131
Bio-layer interferometry 302–305
Biosensor 43, 60, 93, 233, 234, 269–289,
291, 292, 295, 303–305
Biosustainability 109

C

Carotenoids 191, 212, 230, 232, 319–329
CasEMBLR 185–200
Cell factory 99, 269, 286, 307
Chassis 60, 84, 85, 89–92, 95, 103,
211, 212, 334
Clustered regularly interspaced short palindromic repeats
(CRISPR) interference 28
¹³C metabolic flux analysis 333–351
Cofactor specificity 15–26
Cofactor-switch 16, 19
CO₂ fixation 151, 152, 154–158
Computational strain design 63–67, 70
CRiPi 28–32, 35, 36
CRISPR-Cas9 27, 28, 33, 163–181
CRMAGE 28–31, 33–35

D

DNA assembly 41, 131–146, 187,
206, 260, 264
Double-strand break (DSB) 164, 165, 169, 172, 174,
186, 187
Droplet microfluidics 308, 309, 311

E

Ensemble modeling 69, 73
Enzyme reactions 84, 92
Essential genes 27, 28, 30, 36, 165
Evolutionary robustness 39
Expression optimization 40, 42

F

Flippase recombinase 119, 120, 123, 124, 127, 128
Flow-seq 4–6
Fluorescence activated cell sorting (FACS) 6, 11,
233, 308
Flux analysis 68, 335, 338, 346, 351

G

Gene expressions 41, 64, 132, 159, 165,
187, 189, 212, 233, 260, 267, 271, 291, 292
Genetic refactoring 132
Genome engineering 28, 186, 187, 269
Gibson assembly 43, 134, 138, 168,
174, 203–208
Golden Gate assembly 133, 134, 136, 138,
141–145, 214, 219, 225

H

Hierarchical DNA assembly 133
High GC content 163, 164, 203, 205, 207
High-throughput screening 16, 41, 270,
292, 308, 309
Homologous recombination 46, 164, 174,
175, 180, 186, 187, 204, 212, 213, 223, 277,
278, 282, 323
Homology directed repair (HDR) 164, 174

I

Industrial biotechnology 211
In vivo assembly 186, 187, 189–191
Ion mobility spectrometry 256
Isothermal assembly 133, 134, 136, 138,
141, 144, 145, 204

K

Kinetic modeling 63–78

L

Lambda Red recombineering 119, 123–125
Library design 16, 20, 26, 135
Library screening 18

M

Metabolic engineering 15, 39, 59, 64–68, 78, 94,
97, 98, 111, 112, 119, 166, 185, 211, 269, 292,
320, 335, 338, 346, 351
Metabolic modeling 41, 42, 49, 53, 56, 65,
66, 71–73, 99, 110, 334, 339–342, 344
Methanol 150, 151, 153–158, 232, 338
Microbial communities 65, 71, 72
Microbial electrosynthesis (MES) 151, 156, 159
Multi-capillary column (MCC) 236, 237,
240–245, 249, 253–256
Multi-tissue model 65, 73

N

N-degron 28, 29
N-end rule pathway 28, 29
Next-generation sequencing 301
Non-homologous end joining (NHEJ) 164, 165,
169, 171, 173, 177, 180, 225

O

Omics data 230
On-line analysis 231, 235, 237
On-site analysis 229
Operon design 47, 48
Oxidoreductases 16

P

Pathways 4, 15, 27, 39, 70, 84, 98, 132,
151, 165, 211, 286, 292, 320, 339
pCRiPi 31
PCR-ligation 133, 134, 136, 138, 144
Phenotype screening 269, 292
Predictive biology 39, 41, 46
Products 8, 12, 13, 20, 25, 31, 41–43, 46,
49, 52, 53, 55, 60, 63, 64, 67, 69, 74, 76, 85, 88,
89, 94, 95, 98, 99, 107, 109, 111, 114, 115, 121,
124, 126, 132, 136, 138, 139, 141, 144, 145,
159, 163, 166, 171, 172, 174–177, 179, 198,
205, 208, 213, 218–220, 222, 223, 230–233,
253, 267, 278, 279, 281, 284, 300–302, 304,
307, 308, 322, 324, 328
Protein
 engineering 15, 16, 51, 55, 56, 59
 production optimization 261
 secretion 308
 stability 29, 30, 32, 33, 36
Protein interference system (PROTi) 28–35

R

Random mutagenesis 307
Retrosynthetic pathway design 92
RNA aptamers 234, 304

S

Saccharomyces cerevisiae 90, 214, 219, 225,
232, 239, 270, 310, 320
Screening yeast 308
Selection 13, 19, 22, 28–30, 33, 40, 43,
84, 86, 105, 112, 119, 120, 124, 126, 136, 143,
144, 157, 170, 175, 186, 190, 191, 230, 233,
244, 245, 248, 255, 259–268, 270, 271, 276,
278, 283, 291–293, 301, 303, 304, 323
SELEX 291–304
Shine-Dalgarno sequences 3, 4, 7, 46, 263
Small molecules 211, 236, 271, 304
Sporomusa ovata 151, 152, 159
Stoichiometric metabolic models 64, 66, 73
Strain improvement 319
Streptomyces 165, 166, 168, 174,
203–205, 208
Synthetic biology 3, 4, 119, 131,
132, 269, 292, 307, 333, 334
Systems biology 98, 102, 307

T

Techno-economic analysis 98, 109, 115
Transcription factors 233, 234, 270, 271,
276, 277, 286, 292
Translational coupling 259–261, 263, 264, 266

U

Uracil-specific excision reagent (USER)
 cloning 31, 166, 167, 266, 274,
277, 278, 280

V

Versatile genetic assembly system (VEGAS) 212,
214–220, 223, 225
Violacein 214, 219, 225
Volatile metabolites 229–257
Volatile organic compounds (VOCs) 235

W

Whole-plant model 70

Y

Yeast 5, 17, 28, 69, 120, 132, 153,
168, 185, 206, 211, 233, 271, 308, 320
Yeast Golden Gate (yGG) 211–225