Methods in Molecular Biology 1654

Springer Protocols

Michael Kaufmann
Claudia Klinger
Andreas Savelsbergh *Editors*

Functional Genomics

Methods and Protocols

Third Edition

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

Functional Genomics

Methods and Protocols

Third Edition

Edited by

Michael Kaufmann, Claudia Klinger and Andreas Savelsbergh

Chair of Biochemistry and Molecular Medicine, Division of Functional Genomics, Faculty of Health, School of Medicine, Center for Biochemical Research and Education ZBAF, Witten/Herdecke University, Witten, Germany

Humana Press

Editors
Michael Kaufmann
Chair of Biochemistry and Molecular Medicine
Division of Functional Genomics
Faculty of Health
School of Medicine
Center for Biochemical Research and Education ZBAF
Witten/Herdecke University
Witten, Germany

Claudia Klinger
Chair of Biochemistry and Molecular Medicine
Division of Functional Genomics
Faculty of Health
School of Medicine
Center for Biochemical Research and Education ZBAF
Witten/Herdecke University
Witten, Germany

Andreas Savelsbergh
Chair of Biochemistry and Molecular Medicine
Division of Functional Genomics
Faculty of Health
School of Medicine
Center for Biochemical Research and Education ZBAF
Witten/Herdecke University
Witten, Germany

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-7230-2 ISBN 978-1-4939-7231-9 (eBook)
DOI 10.1007/978-1-4939-7231-9

Library of Congress Control Number: 2017949390

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

Functional genomics is characterized as a field in life science making use of high-throughput experimental methods combined with exhaustive bioinformatic analysis in order to discover complex relationships between genotype and phenotype on a genome-wide scale. The field came into being about three decades ago with the development of the first microarray technologies at that time mainly dealing with expression profiling which consequently was the major focus of the first edition of the book, *Functional Genomics: Methods and Protocols*. With our second edition, we already extended the scope by including bioinformatics as well as protein and metabolite analysis. Again, this third edition opens with a chapter on bioinformatic procedures suitable to make both structural and functional predictions about RNA and proteins. Although also representing some functional aspects of the genome, metabolites and the metabolome were somewhat neglected in science recently. For that reason, we did not include metabolite analysis in this third edition.

Any high-throughput experiment typically produces vast amounts of data and is therefore, at least in part, an explorative rather than a hypothesis-driven scientific approach. We believe that the predominantly explorative -omics era is slowly beginning to be replaced by more hypothesis-driven concepts. This will lead to many functional aspects of the genome getting specifically settled. Even more, the genome just became a subject of manipulation as genome editing via CRISPR/CAS impressively demonstrates, and thus along with other sophisticated methods it stands as a new trend. To take into account this tendency, we changed the division of the biochemical chapters and included the part: "From Genotype to Phenotype."

During the last years, it became more and more clear that most of the information stored in the eukaryotic genome and most of the energy spent in eukaryotic metabolism are utilized for regulation. Junk DNA no longer seems to have anything to do with junk, as clearly can be seen by the fact that although less than 5% of the human genome contain protein-coding information, more than 80% are transcribed resulting in an impressive excess of noncoding RNA over mRNA. These transcripts can by definition be attributed to regulatory functions. In contrast to resolving classical almost linearly organized biochemical pathways, investigating regulatory networks such as RNA interference will be much more complex and, as we feel, will even be the main challenge of functional genomics in the future.

Our thanks go to all authors for contributing their carefully arranged manuscripts and especially for their great patience with our painstaking editing process. May the reader profit from the protocols all described as accurately as possible in order to keep experimental failures, and thus frustration, to a minimum.

Witten, Germany

Michael Kaufmann
Claudia Klinger
Andreas Savelsbergh

Contents

Preface . . . . . *v*
Contributors . . . . . *xi*

PART I BIOINFORMATICS

1 Predicting RNA Structure with Vfold 3
Chenhan Zhao, Xiaojun Xu, and Shi-Jie Chen

2 RNA Function Prediction 17
Yongsheng Li, Juan Xu, Tingting Shao, Yunpeng Zhang, Hong Chen, and Xia Li

3 Computational Prediction of Novel miRNAs from Genome-Wide Data 29
Georgina Stegmayer, Cristian Yones, Laura Kamenetzky, Natalia Macchiaroli, and Diego H. Milone

4 Protein Structure Modeling with MODELLER 39
Benjamin Webb and Andrej Sali

5 Protein Function Prediction 55
Leonardo Magalhães Cruz, Sheyla Trefflich, Vinícius Almir Weiss, and Mauro Antônio Alves Castro

PART II DNA ANALYSIS

6 Capturing Three-Dimensional Genome Organization in Individual Cells by Single-Cell Hi-C 79
Takashi Nagano, Steven W. Wingett, and Peter Fraser

7 Genome-Wide Cell Type-Specific Mapping of In Vivo Chromatin Protein Binding Using an FLP-Inducible DamID System in *Drosophila* 99
Alexey V. Pindyurin

8 DNA Methylation Profiling Using Long-Read Single Molecule Real-Time Bisulfite Sequencing (SMRT-BS) 125
Yao Yang and Stuart A. Scott

9 Copy Number Variation Analysis by Droplet Digital PCR 135
Suvi K. Härmälä, Robert Butcher, and Chrissy H. Roberts

10 MicroScale Thermophoresis: A Rapid and Precise Method to Quantify Protein–Nucleic Acid Interactions in Solution 151
Adrian Michael Mueller, Dennis Breitsprecher, Stefan Duhr, Philipp Baaske, Thomas Schubert, and Gernot Längst

11 Establishment of the CRISPR/Cas9 System for Targeted Gene Disruption and Gene Tagging 165
Eric Ehrke-Schulz, Maren Schiwon, Claudia Hagedorn, and Anja Ehrhardt

Part III RNA Analysis

12 Holistic and Affordable Analyses of MicroRNA Expression Profiles Using Tagged cDNA Libraries and a Multiplex Sequencing Strategy 179
Patrick P. Weil, Yan Jaszczyszyn, Anne Baroin-Tourancheau, Jan Postberg, and Laurence Amar

13 MicroRNA Expression Analysis Using Small RNA Sequencing Discovery and RT-qPCR-Based Validation 197
Alan Van Goethem, Pieter Mestdagh, Tom Van Maerken, and Jo Vandesompele

14 Using FirePlex™ Particle Technology for Multiplex MicroRNA Profiling Without RNA Purification 209
Michael R. Tackett and Izzuddin Diwan

15 Multiplex Real-Time PCR Using Encoded Microparticles for MicroRNA Profiling 221
Seungwon Jung and Sang Kyung Kim

16 Optimized Whole Transcriptome Profiling of Motor Axons 231
Lena Saal-Bauernschubert, Michael Briese, and Michael Sendtner

Part IV Protein Analysis

17 2D-DIGE in Proteomics 245
Matias Pasquali, Tommaso Serchi, Sebastien Planchon, and Jenny Renaut

18 STAGE-Diging in Proteomics 255
Paolo Soffientini and Angela Bachi

19 Protein Arrays I: Antibody Arrays 261
Yulin Yuan, Zuan-Tao Lin, Hongting Wang, Xia Hong, Mikala Heon, and Tianfu Wu

20 Protein Arrays II: Antigen Arrays 271
Yulin Yuan, Hongting Wang, Zuan-Tao Lin, Xia Hong, Mikala Heon, and Tianfu Wu

21 Protein Arrays III: Reverse-Phase Protein Arrays 279
Yulin Yuan, Xia Hong, Zuan-Tao Lin, Hongting Wang, Mikala Heon, and Tianfu Wu

22 Isolation of Exosomes for the Purpose of Protein Cargo Analysis with the Use of Mass Spectrometry 291
Monika Pietrowska, Sonja Funk, Marta Gawin, Łukasz Marczak, Agata Abramowicz, Piotr Widłak, and Theresa Whiteside

Part V From Genotype to Phenotype

23 Virus-Induced Gene Silencing (VIGS) and Foreign Gene Expression in *Pisum sativum* L. Using the "One-Step" *Bean pod mottle virus* (BPMV) Viral Vector 311
Chouaïb Meziadi, Sophie Blanchet, Valérie Geffroy, and Stéphanie Pflieger

24 Re-expressing Epigenetically Silenced Genes by Inducing DNA Demethylation Through Targeting of Ten-Eleven Translocation 2 to Any Given Genomic Locus 321
Julio Cesar Rendón, David Cano-Rodríguez, and Marianne G. Rots

25 Knockdown of Rice microRNA166 by Short Tandem Target Mimic (STTM) 337
Sachin Teotia, Dabing Zhang, and Guiliang Tang

26 RNAi-Mediated Knockdown of Protein Expression 351
Volker Baumann, Cornelia Lorenzer, Michael Thell, Anna-Maria Winkler, and Johannes Winkler

27 Engineered Zinc Finger DNA-Binding Domains: Synthesis, Assessment of DNA-Binding Affinity, and Direct Protein Delivery to Mammalian Cells 361
Mir A. Hossain, Isaac J. Knudson, Shaleen Thakur, Yong Shen, Jared R. Stees, Joeva J. Barrow, and Jörg Bungert

28 Production, Purification, and Titration of First-Generation Adenovirus Vectors 377
Ramona F. Kratzer and Florian Kreppel

Index 389

Contributors

Agata Abramowicz • *Center for Translational Research and Molecular Biology of Cancer, Maria Sklodowska—Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch, Gliwice, Poland*

Laurence Amar • *Institut des Neurosciences Paris-Saclay, Université Paris-Sud, CNRS, UMR 9197, Université Paris-Saclay, Orsay, France*

Philipp Baaske • *NanoTemper Technologies GmbH, Munich, Germany*

Angela Bachi • *IFOM, FIRC Institute of Molecular Oncology, Milan, Italy*

Anne Baroin-Tourancheau • *Institut des Neurosciences Paris-Saclay, Université Paris-Sud, CNRS, UMR 9197, Université Paris-Saclay, Orsay, France*

Joeva J. Barrow • *Department of Cancer Biology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA; Department of Cell Biology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA*

Volker Baumann • *Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria*

Sophie Blanchet • *Institute of Plant Sciences-Paris Saclay (IPS2), CNRS, INRA, Université Paris-Sud, Université d'Evry, Université Paris Saclay, Gif sur Yvette CEDEX, France; Institute of Plant Sciences-Paris Saclay (IPS2), Université Paris Diderot, Sorbonne Paris-Cité, Gif sur Yvette CEDEX, France*

Dennis Breitsprecher • *NanoTemper Technologies GmbH, Munich, Germany*

Michael Briese • *Institute for Clinical Neurobiology, University Hospital Wuerzburg, Wuerzburg, Germany*

Jörg Bungert • *Department of Biochemistry and Molecular Biology, College of Medicine, UF Health Cancer Center, Genetics Institute, University of Florida, Gainesville, FL, USA*

Robert Butcher • *Department of Clinical Research, London School of Hygiene & Tropical Medicine, London, UK*

David Cano-Rodríguez • *Epigenetic Editing Research Group, Department of Pathology and Medical Biology, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands*

Mauro Antônio Alves Castro • *Sector of Professional and Technological Education, Federal University of Paraná (UFPR), Curitiba, PR, Brazil*

Shi-Jie Chen • *Department of Physics, Informatics Institute, University of Missouri, Columbia, MO, USA*

Hong Chen • *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China*

Leonardo Magalhães Cruz • *Department of Biochemistry and Molecular Biology, Federal University of Paraná (UFPR), Curitiba, PR, Brazil; Sector of Professional and Technological Education, Federal University of Paraná (UFPR), Curitiba, PR, Brazil*

Izzuddin Diwan • *Abcam Inc., Cambridge, MA, USA*

Stefan Duhr • *NanoTemper Technologies GmbH, Munich, Germany*

Anja Ehrhardt • *Chair of Virology and Microbiology, Faculty of Health, School of Medicine, Center for Biomedical Research and Education ZBAF, Witten/Herdecke University, Witten, Germany*

Eric Ehrke-Schulz • *Chair of Virology and Microbiology, Faculty of Health, School of Medicine, Center for Biomedical Research and Education ZBAF, Witten/Herdecke University, Witten, Germany*
Peter Fraser • *Nuclear Dynamics Programme, The Babraham Institute, Cambridge, UK*
Sonja Funk • *Department of Pathology, University of Pittsburgh, Pittsburgh, PA, USA; Department of Otolaryngology, University of Duisburg-Essen, Essen, Germany*
Marta Gawin • *Center for Translational Research and Molecular Biology of Cancer, Maria Sklodowska—Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch, Gliwice, Poland*
Valérie Geffroy • *Institute of Plant Sciences-Paris Saclay (IPS2), CNRS, INRA, Université Paris-Sud, Université d'Evry, Université Paris Saclay, Gif sur Yvette CEDEX, France; Institute of Plant Sciences-Paris Saclay (IPS2), Université Paris Diderot, Sorbonne Paris-Cité, Gif sur Yvette CEDEX, France*
Suvi K. Härmälä • *MRC International Nutrition Group, London School of Hygiene & Tropical Medicine, London, UK*
Claudia Hagedorn • *Chair of Biochemistry and Molecular Medicine, Faculty of Health, School of Medicine, Center for Biomedical Research and Education ZBAF, Witten/ Herdecke University, Witten, Germany*
Mikala Heon • *Department of Biomedical Engineering, University of Houston, Houston, TX, USA*
Xia Hong • *Department of Biomedical Engineering, University of Houston, Houston, TX, USA; Department of Nursing, Fujian Health College, Fuzhou, Fujian, China*
Mir A. Hossain • *Department of Biochemistry and Molecular Biology, College of Medicine, UF Health Cancer Center, Genetics Institute, University of Florida, Gainesville, FL, USA*
Yan Jaszczyszyn • *Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette, France*
Seungwon Jung • *Center for BioMicrosystems, Brain Science Institute, Korea Institute of Science and Technology, Seoul, South Korea*
Laura Kamenetzky • *Instituto de Investigaciones en Microbiología y Parasitología Médica, IMPaM, CONICET-UBA, Buenos Aires, Argentina*
Sang Kyung Kim • *Center for BioMicrosystems, Brain Science Institute, Korea Institute of Science and Technology, Seoul, South Korea; Department of Biomedical Engineering, Korea University of Science and Technology, Daejeon, South Korea*
Isaac J. Knudson • *Department of Biochemistry and Molecular Biology, College of Medicine, UF Health Cancer Center, Genetics Institute, University of Florida, Gainesville, FL, USA*
Ramona F. Kratzer • *Department of Gene Therapy, Ulm University, Ulm, Germany*
Florian Kreppel • *Chair of Biochemistry and Molecular Medicine, Faculty of Health, School of Medicine, Center for Biomedical Research and Education ZBAF, Witten/ Herdecke University, Witten, Germany*
Gernot Längst • *Biochemistry III, University of Regensburg, Regensburg, Germany*
Yongsheng Li • *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China*
Xia Li • *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China*
Zuan-Tao Lin • *Department of Biomedical Engineering, University of Houston, Houston, TX, USA*
Cornelia Lorenzer • *Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria*

Natalia Macchiaroli • *Instituto de Investigaciones en Microbiología y Parasitología Médica, IMPaM, CONICET-UBA, Buenos Aires, Argentina*
Łukasz Marczak • *Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland*
Pieter Mestdagh • *Center for Medical Genetics, Ghent University, Ghent, Belgium*
Chouaïb Meziadi • *Institute of Plant Sciences-Paris Saclay (IPS2), CNRS, INRA, Université Paris-Sud, Université d'Evry, Université Paris Saclay, Gif sur Yvette CEDEX, France; Institute of Plant Sciences-Paris Saclay (IPS2), Université Paris Diderot, Sorbonne Paris-Cité, Gif sur Yvette CEDEX, France*
Diego H. Milone • *Research Institute for Signals, Systems and Computational Intelligence, sinc(i), CONICET-UNL, Santa Fe, Argentina*
Adrian Michael Mueller • *Biochemistry III, University of Regensburg, Regensburg, Germany*
Takashi Nagano • *Nuclear Dynamics Programme, The Babraham Institute, Cambridge, UK*
Matias Pasquali • *Luxembourg Institute of Science and Technology, Belvaux, Luxembourg; DEFENS, University of Milan, Milan, Italy*
Stéphanie Pflieger • *Institute of Plant Sciences-Paris Saclay (IPS2), CNRS, INRA, Université Paris-Sud, Université d'Evry, Université Paris Saclay, Gif sur Yvette CEDEX, France; Institute of Plant Sciences-Paris Saclay (IPS2), Université Paris Diderot, Sorbonne Paris-Cité, Gif sur Yvette CEDEX, France*
Monika Pietrowska • *Center for Translational Research and Molecular Biology of Cancer, Maria Sklodowska—Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch, Gliwice, Poland*
Alexey V. Pindyurin • *Laboratory of Cell Division, Institute of Molecular and Cellular Biology SB RAS, Novosibirsk, Russia*
Sebastien Planchon • *Luxembourg Institute of Science and Technology, Belvaux, Luxembourg*
Jan Postberg • *Department of Paediatrics, HELIOS Medical Centre Wuppertal, Centre for Clinical and Translational Research (CCTR), Witten/Herdecke University Hospital, Centre for Biomedical Education and Research (ZBAF), Witten, Germany*
Jenny Renaut • *Luxembourg Institute of Science and Technology, Belvaux, Luxembourg*
Julio Cesar Rendón • *Epigenetic Editing Research Group, Department of Pathology and Medical Biology, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands*
Chrissy H. Roberts • *Department of Clinical Research, London School of Hygiene & Tropical Medicine, London, UK*
Marianne G. Rots • *Epigenetic Editing Research Group, Department of Pathology and Medical Biology, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands*
Lena Saal-Bauernschubert • *Institute for Clinical Neurobiology, University Hospital Wuerzburg, Wuerzburg, Germany*
Andrej Sali • *Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences (QB3), University of California San Francisco, San Francisco, CA, USA*
Maren Schiwon • *Chair of Virology and Microbiology, Faculty of Health, School of Medicine, Center for Biomedical Research and Education ZBAF, Witten/Herdecke University, Witten, Germany*

THOMAS SCHUBERT • *2bind GmbH, Regensburg, Germany*
STUART A. SCOTT • *Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA*
MICHAEL SENDTNER • *Institute for Clinical Neurobiology, University Hospital Wuerzburg, Wuerzburg, Germany*
TOMMASO SERCHI • *Luxembourg Institute of Science and Technology, Belvaux, Luxembourg*
TINGTING SHAO • *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China*
YONG SHEN • *Department of Biochemistry and Molecular Biology, College of Medicine, UF Health Cancer Center, Genetics Institute, University of Florida, Gainesville, FL, USA*
PAOLO SOFFIENTINI • *IFOM, FIRC Institute of Molecular Oncology, Milan, Italy*
JARED R. STEES • *Department of Biochemistry and Molecular Biology, College of Medicine, UF Health Cancer Center, Genetics Institute, University of Florida, Gainesville, FL, USA*
GEORGINA STEGMAYER • *Research Institute for Signals, Systems and Computational Intelligence, sinc(i), CONICET-UNL, Santa Fe, Argentina*
MICHAEL R. TACKETT • *Abcam Inc., Cambridge, MA, USA*
GUILIANG TANG • *Department of Biological Sciences, Michigan Technological University, Houghton, MI, USA*
SACHIN TEOTIA • *Department of Biological Sciences, Michigan Technological University, Houghton, MI, USA*
SHALEEN THAKUR • *Department of Biochemistry and Molecular Biology, College of Medicine, UF Health Cancer Center, Genetics Institute, University of Florida, Gainesville, FL, USA*
MICHAEL THELL • *Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria*
SHEYLA TREFFLICH • *Sector of Professional and Technological Education, Federal University of Paraná (UFPR), Curitiba, PR, Brazil*
ALAN VAN GOETHEM • *Center for Medical Genetics, Ghent University, Ghent, Belgium*
TOM VAN MAERKEN • *Center for Medical Genetics, Ghent University, Ghent, Belgium*
JO VANDESOMPELE • *Center for Medical Genetics, Ghent University, Ghent, Belgium*
HONGTING WANG • *Department of Biomedical Engineering, University of Houston, Houston, TX, USA; National Pharmacology Laboratory of Chinese Medicine, College of Basic Medical Sciences, Wannan Medical College, Wuhu, Anhui, China*
BENJAMIN WEBB • *Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences (QB3), University of California San Francisco, San Francisco, CA, USA*
PATRICK P. WEIL • *Department of Paediatrics, HELIOS Medical Centre Wuppertal, Centre for Clinical and Translational Research (CCTR), Witten/Herdecke University Hospital, Centre for Biomedical Education and Research (ZBAF), Witten, Germany*
VINÍCIUS ALMIR WEISS • *Sector of Professional and Technological Education, Federal University of Paraná (UFPR), Curitiba, PR, Brazil*
THERESA WHITESIDE • *Department of Pathology, University of Pittsburgh, Pittsburgh, PA, USA*
PIOTR WIDŁAK • *Center for Translational Research and Molecular Biology of Cancer, Maria Sklodowska—Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch, Gliwice, Poland*
STEVEN W. WINGETT • *Nuclear Dynamics Programme, The Babraham Institute, Cambridge, UK*

Anna-Maria Winkler • *Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria*
Johannes Winkler • *Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria; Department of Cardiology, Medical University of Vienna, Vienna, Austria*
Tianfu Wu • *Department of Biomedical Engineering, University of Houston, Houston, TX, USA*
Xiaojun Xu • *Department of Physics, Informatics Institute, University of Missouri, Columbia, MO, USA*
Juan Xu • *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China*
Yao Yang • *Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA*
Cristian Yones • *Research Institute for Signals, Systems and Computational Intelligence, sinc(i), CONICET-UNL, Santa Fe, Argentina*
Yulin Yuan • *Department of Biomedical Engineering, University of Houston, Houston, TX, USA; Department of Clinical Laboratory, The People's Hospital of Guangxi Zhuang Autonomous Region, Nanning, Guangxi, China*
Dabing Zhang • *School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China*
Yunpeng Zhang • *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China*
Chenhan Zhao • *Department of Physics, Informatics Institute, University of Missouri, Columbia, MO, USA*

Part I

Bioinformatics

Chapter 1

Predicting RNA Structure with Vfold

Chenhan Zhao, Xiaojun Xu, and Shi-Jie Chen

Abstract

In order to carry out biological functions, RNA molecules must fold into specific three-dimensional (3D) structures. Current experimental methods to determine RNA 3D structures are expensive and time consuming. With the recent advances in computational biology, RNA structure prediction is becoming increasingly reliable. This chapter describes a recently developed RNA structure prediction software, Vfold, a virtual bond-based RNA folding model. The main features of Vfold are the physics-based loop free energy calculations for various RNA structure motifs and a template-based assembly method for RNA 3D structure prediction. For illustration, we use the *yybP-ykoY* Orphan riboswitch as an example to show the implementation of the Vfold model in RNA structure prediction from the sequence.

Key words RNA folding, Vfold model, Loop entropy, Template assembly

1 Introduction

Current experimental methods, such as X-ray crystallography [1], NMR [2], and electron microscopy [3], can determine RNA 3D structures in high or low resolutions. However, using experimental methods to determine RNA structures can be expensive and time consuming. With the rapid advances of RNA sequencing technology [4], experimental methods may not catch up the demands for high resolution RNA 3D structures. Therefore, computational structure prediction becomes a highly needed tool for RNA biology.

An RNA structure can be described at 2D and 3D levels. A 2D structure is defined by the base pairs contained in the structure, which provides structural constraints for 3D structure folding. Current RNA 2D structure prediction algorithms can be classified into two major categories [5–10]: sequence alignment-based methods and free energy-based methods. In general, sequence alignment software, such as Dynalign [11], gives reliable 2D structures if homologous sequences are available. However, many alternative structures, which may not be predicted by the comparative

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_1,

sequence analysis method, can also be functionally important. For example, riboswitches undergo a conformational change in response to binding of a regulatory molecule [12]. Free energy-based methods, such as Mfold [13], RNAstructure [14], and RNAfold [15], calculate the free energies for an ensemble of structures and find the minimum free energy structure or the most probable (average) structure. One of the key ingredients of these methods is the availability of thermodynamic parameters for loops and helices. The thermodynamic parameters for helices and simple loops (i.e., small-size hairpin, internal/bulge loops) have been determined systematically and compiled as the Turner's parameters [16]. However, free energy parameters of other more complicated loops remain unknown and need to be determined through a computational model.

Knowing RNA 2D structure is not sufficient to obtain high resolution 3D structure. In general, a 2D structure can correspond to a large number of 3D structures due to the multiplicity of flexible loop conformations. We still need methods to model the structures of the unpaired nucleotides and the relative orientations of helices. There are many different ways to predict RNA 3D structures from given 2D structure. For example, one such method is to use knowledge-based force field and predict RNA 3D structures from coarse-grained discrete molecular dynamics (DMD) simulations [17–20]. Here the coarse-grained representation for RNA conformations can dramatically decrease the number of freedoms of an RNA system, and thus increase the completeness of the conformational sampling. One of the major issues in the simulations is that the sampled conformations often remain close to the initial starting model, which requires the use of various special simulation techniques to achieve effective sampling of conformational space. One of the attempts to circumvent this problem is to use template-based structure prediction algorithm [21–24]. For the template-based approaches, one of the common limitations is the limited degree of divergence of the template library. Given the limited number of known RNA structures, structural motif templates with the required high sequence similarity are difficult to attain. The lack of reliable structural motifs for many loops and junctions has greatly hampered our effort for successful 3D structure prediction. Nevertheless, as more and more RNA structures are experimentally determined, we can realistically expect the continuous improvements in the accuracy of structure prediction using template-based prediction algorithms.

The recently developed Vfold model [23–26] is a free energy-based RNA folding model to predict RNA structures and thermodynamic stabilities from the sequence. Compared with other RNA structure prediction software [13–15, 22], Vfold uses a coarse-grained representation [24, 25] for RNA conformations. The model enumerates all the possible loop conformations in 3D

space to calculate loop entropy and free energy parameters. For the 3D structure prediction, Vfold uses template-based method to assemble RNA 3D structures from motifs. In this chapter, we illustrate the application of the Vfold software/web server [24] in RNA structure prediction.

2 Algorithms

2.1 Computation of Loop Entropies and Prediction of 2D Structure (Vfold2D)

Vfold model uses two virtual bonds (P-C4′ and C4′-P) per nucleotide to represent RNA backbone configurations (*see* Fig. 1). By enumerating all the possible virtual-bond conformations in the 3D space (*see* **Note 1**), Vfold estimates the loop entropy parameters from the probability of loop closure [24, 25]. The model has the advantage of accounting for chain connectivity, excluded volume (between loops and helices), and the completeness of virtual-bonded loop conformational ensemble. Vfold2D [24] is a free energy-based model that predicts RNA 2D structures using the above Vfold-derived entropy and free energy parameters.

Here, we use the pseudoknotted loop structure to illustrate the calculation for the Vfold entropy parameters. A pseudoknotted motif, as shown in Fig. 1c, consists of two helical stems and three loops. The relative orientation of the two helices can be configured by the 3D conformation of the L_2 loop. Loops and helices can be correlated. For example, the loop conformations are constrained by the loop–helix volume exclusion, and the helix orientations can be

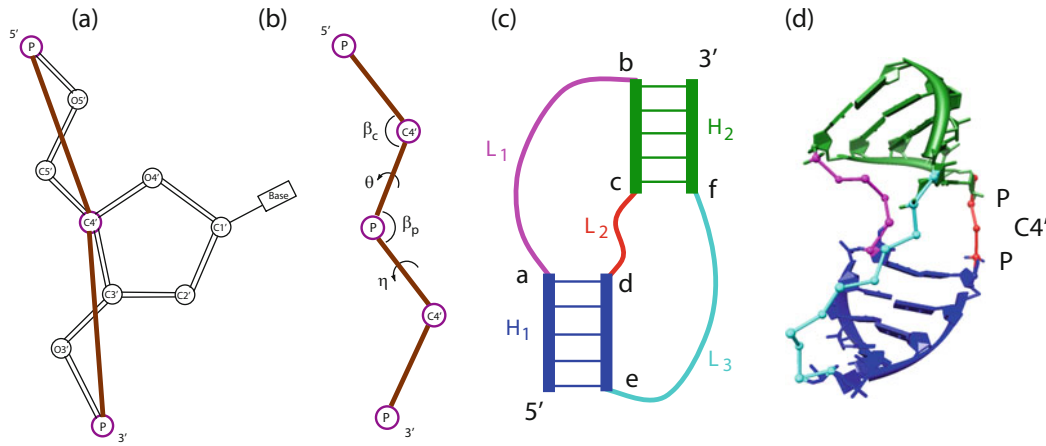


Fig. 1 The Vfold model uses two bonds (P-C4′ and C4′-P) to represent each nucleotide and computes loop entropies by sampling virtual-bonded conformations in 3D space. (**a**) Virtual-bonded representation of an RNA nucleotide. (**b**) The bond angles (β_c, β_p) and the torsional angles (θ, η) for the virtual bonds. Vfold enumerates RNA backbone conformations on a diamond lattice with bond length of 3.9 Å, bond angle of $\sim 109.5^\circ$ and three equiprobable torsional angles (60°, 180°, 300°). (**c**) A schematic diagram for a pseudoknotted loop. (**d**) A virtual-bonded pseudoknotted loop structure with all-atom helices

determined from the coordinates of the nucleotides (*a*, *b*, *c*, *d*, *e*, and *f* in Fig. 1c) in the loop. Therefore, the free energy change, especially the entropic decrease, for the formation of the pseudoknotted loop structure depends not only on the lengths of the single-stranded loops L_1, L_2, and L_3 but also on the lengths of the helices H_1 and H_2. Since the two virtual bonds per nucleotide used in Vfold model describe only the backbone structures, the Vfold-derived loop entropy parameters do not account for the sequence dependence per se (*see* **Note 2**). However, by explicitly enumerating the sequence-dependent intraloop structures (such as mismatches) for the loops, the Vfold2D model can (partially) account for the sequence dependence of the loop free energy. The computation of the loop entropy parameters in the Vfold2D model involves the following three steps.

1. We sample helix configurations by enumerating the virtual-bond conformations of loop L_2. The connection between the A-form helices and the discrete loop conformations is realized through an iterative optimized algorithm [27]. Helices are modeled as the all-atom A-form [28] helix structures (*see* **Note 3**).
2. For each helix orientation, with the given (*a*, *b*) of the starting and ending nucleotides for loop L_1 and (*e*, *f*) of the starting and ending nucleotides for loop L_3, we sample loop conformations as self-avoiding walks of the virtual bonds on the diamond lattice to sample loops/junctions 3D conformations (*see* **Note 4**).
3. We estimate the loop entropy parameter as the logarithm of the probability of loop formation: $\Delta S_{\text{loop}} = k_{\text{B}} \ln(\Omega_{\text{loop}}/\Omega_{\text{coil}})$ (*see* **Note 2**). Here Ω_{loop} and Ω_{coil} are the conformational counts of the loop and the coil structures, respectively, and k_B is the Boltzmann constant.

With the Vfold-derived loop entropy parameters [25, 29–32] and the experimentally determined base stacking thermodynamic parameters [16], Vfold2D [24] gives the free energy for each 2D structure and hence predicts the minimum free energy structure and the possible alternative metastable structures.

2.2 VfoldMTF: A Database of RNA 3D Motifs

To predict the 3D structure from the 2D structure using the knowledge about the known structures, we need to construct a database for all the known structural motifs. We have complied a database "VfoldMTF" (see http://rna.physics.missouri.edu/vfoldMTF/) for the 3D structural motifs, including hairpin loops, internal/bulge loops, *N*-way junctions ($2 < N < 8$), H-type pseudoknots, hairpin/hairpin kissing motifs, and two-way/hairpin kissing motifs (*see* Fig. 2). The database shows the sequence of each motif as well as the PDB IDs of all the PDB entries that

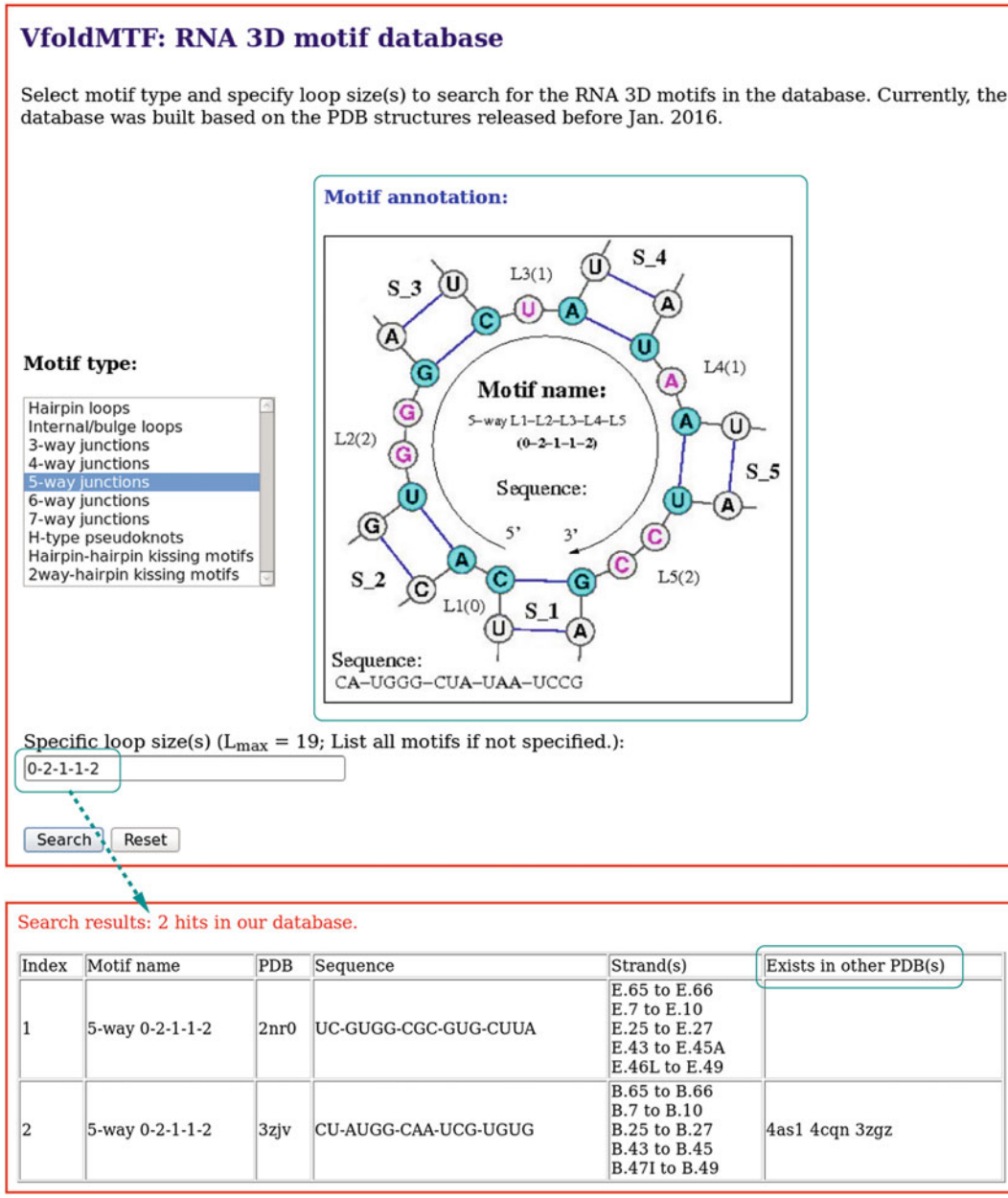


Fig. 2 Snapshot of the VfoldMTF database. "Motif annotation" denotes the definition of the motifs. Users can search for motifs (with given loop sizes, optional) within the database. The output of VfoldMTF gives the information about the sequences and the strand(s), as well as the PDB id(s). For example, the five-way junction with loop sizes 0-2-1-1-2 in the structure 3zjv (PDB id) can also be found in 4as1, 4cqn, and 3zgz. The information may be helpful for structure-function analysis

contain the motif. Currently, the raw database for the 3D motifs is built from 2626 known RNA 3D structures, including all the structures involving RNA (except for RNA/DNA hybrids). With the increasing number of PDB entries, the VfoldMTF database will be continuously updated.

Here are the methods to extract motif templates from known RNA 3D structures and build a non-redundant 3D motif database:

1. For a given RNA 3D structure (*see* **Note 5**), extract the A-form helices (*see* **Note 6**).
2. Determine the corresponding 2D structure for the given 3D structure based on the helices and loops.
3. Identify all the non-helix 2D structure motifs (such as hairpin loop, internal loop, three-way junction, and H-type pseudoknot).
4. Remove the redundant templates for those with RMSD (Root Mean Square Deviation) < 1.5 Å for the same motif type and same sequence.
5. Collect all the non-redundant templates to construct an RNA 3D motif database.

This new database distinguishes itself from other database [33–36] in the treatment of mismatches and other non-canonical interactions. For example, we consider nucleotides involved in the non-canonical base pairs (mismatches) as unpaired nucleotides. In addition, we classify motifs according to loop types (such as hairpin loops, *N*-way junctions, and hairpin–hairpin kissing motifs) and their sizes instead of the type of intra-loop interactions (*see* Fig. 2). The database can be used not only for the motif template-assembly method for RNA 3D structure prediction, but also for the analysis of structure–function relationships.

2.3 3D Structure Prediction Through Motif-Template Assembly (Vfold3D)

Vfold3D, a package of Vfold for RNA 3D structure prediction, uses the template-based method to build RNA 3D structures [23, 24]. Compared with other similar approaches, such as FARNA [21] and MC-Sym [22], Vfold3D uses motif-based templates instead of fragment-based templates. The method can account for the intra-motif interactions (*see* **Note 7**). Predicting the 3D structures from the sequence and the 2D structure (base pairs) involves the following steps.

1. Vfold3D first extracts motifs (such as helices (*see* **Note 8**), hairpin loops, internal/bulge loops, and *N*-way junctions) from the given 2D structure.
2. Helices are modeled as the A-form virtual-bonded helix structures.

3. For the non-helix motifs, Vfold3D searches for the best templates from the VfoldMTF database to identify the appropriate template structures. The search criteria are based on the size (first) and sequence (second) matches (*see* **Note 9**). If necessary, this step may involve sequence replacement and base pair(s) insertion or deletion in order to match the templates in the database.
4. Vfold3D assembles the helix and loop 3D virtual-bonded structures to construct the 3D scaffold of the whole RNA.
5. Vfold3D adds all atoms to the virtual-bonded structure. For nucleotides in each helix, atoms are added according to the A-form helix atomic structure. The all-atom nucleotides in loops are generated by adding atoms according to the templates for base configurations.
6. The assembled all-atom structures are refined by the all-atom energy minimization (*see* **Note 10**).

3 Methods

To predict RNA 3D structures, Vfold first predicts the 2D structures from the sequence through the Vfold2D package. Using the 2D structures as constraint, the model then predicts the corresponding 3D structures from the Vfold3D package. The Vfold web server is freely accessible at http://rna.physics.missouri.edu.

3.1 To Predict RNA 2D Structures with Vfold2D

Vfold2D uses the Vfold-derived pre-tabulated loop entropy parameters [25, 29–32] (*see* **Note 11**) to evaluate loop stability for each sampled structure. Currently, the Vfold2D server can predict RNA 2D structures for (a) secondary (non-cross-linked) structure ensemble of sequence length less than 300 nucleotides and (b) H-type pseudoknotted structure ensemble of length less than 150 nucleotides, due to the long computational time. In addition to the lowest free energy structure, Vfold2D can also predict alternative structures.

1. Visit the Vfold2D server at (http://rna.physics.missouri.edu/vfold2D).
2. The input of Vfold2D (*see* Fig. 3a) contains RNA sequence (A, a,U,u,G,g,C,c letters only), temperature in Celsius and the choice of the energy parameters for base stacks (including mismatched stacks), which can be either from the Turner's parameters (04 version) [16] or the MFOLD (2.3 version) [13] (*see* **Note 12**).
3. The computational time of Vfold2D depends on the length of input sequence (*see* **Note 13**). We recommend users to provide

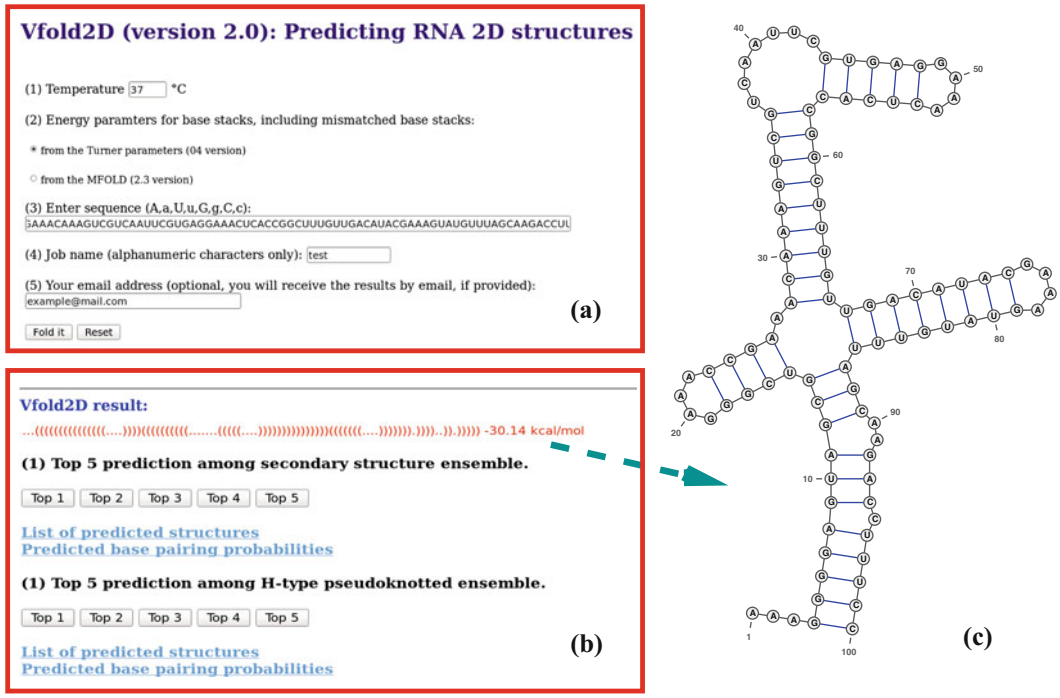


Fig. 3 An example of the Vfold2D prediction. (**a**) Input interface of Vfold2D server. Users have options to input sequence, temperature and choose base stack parameters for helices. (**b**) Output interface of Vfold2D server. Depending on the length of the input sequence, there are up to two sets of predictions (secondary and H-type pseudoknotted structures, respectively), with a list of predicted structures and the base pairing probabilities, included in the result page. For clarity, the 2D structure of the top 1 prediction is shown in (**c**)

emails before submission, so the Vfold2D predictions can be delivered when the computation is finished.

4. Once the calculation is finished, users can retrieve the Vfold2D predictions either from the job-specific result page, as shown in Fig. 3, or from the email. Depending on the length of the input sequence, Vfold2D outputs up to two sets of predictions (for secondary and H-type pseudoknotted structures, respectively), with a list of predicted alternative structures and the base pairing probabilities (*see* **Note 14**).
5. Users can also check the top 5 (if available) 2D structures, plotted by VARNA [37], for each set of the predicted structures on the result page (*see* Fig. 3).

We use the *yybP-ykoY* orphan riboswitch [38] (PDB:4y1i) as an example to illustrate the Vfold2D prediction. As shown in Fig. 3a, the input of Vfold2D is the 100-nucleotide long sequence (*see* **Note 15**), with the temperature of 37 °C and the helix energy parameters from the Turner's (04 version) parameters. Vfold2D outputs two sets of predicted results. The most probable 2D structure is the top 1 prediction from the secondary structure ensemble,

as shown in Fig. 3c, which has the free energy of −30.14 kcal/mol. Compared with the native 2D structure [38] (*see* **Note 16**), Vfold2D correctly predicts 28 of 33 (84.8%) canonical base pairs in the native one.

3.2 To Predict RNA 3D Structures with Vfold3D

With the predicted 2D structure by Vfold2D (*see* **Note 17**), users can predict the 3D structures using the Vfold3D web server. Due to the limited divergence of the current VfoldMTF database, the current version of Vfold3D can only predict RNA 3D structures with hairpin loops, junctions, and limited number of pseudo-knotted motifs.

1. Vfold3D web server is accessible at (http://rna.physics.missouri.edu/vfold3D).
2. The input of Vfold3D is the RNA sequence and the corresponding 2D structure (base pair information) in dot-bracket format (*see* Fig. 4a).
3. Users can also leave their email addresses to receive the Vfold3D results through email.
4. Vfold3D may predict multiple all-atom 3D structures if multiple optimal templates are available in the database.

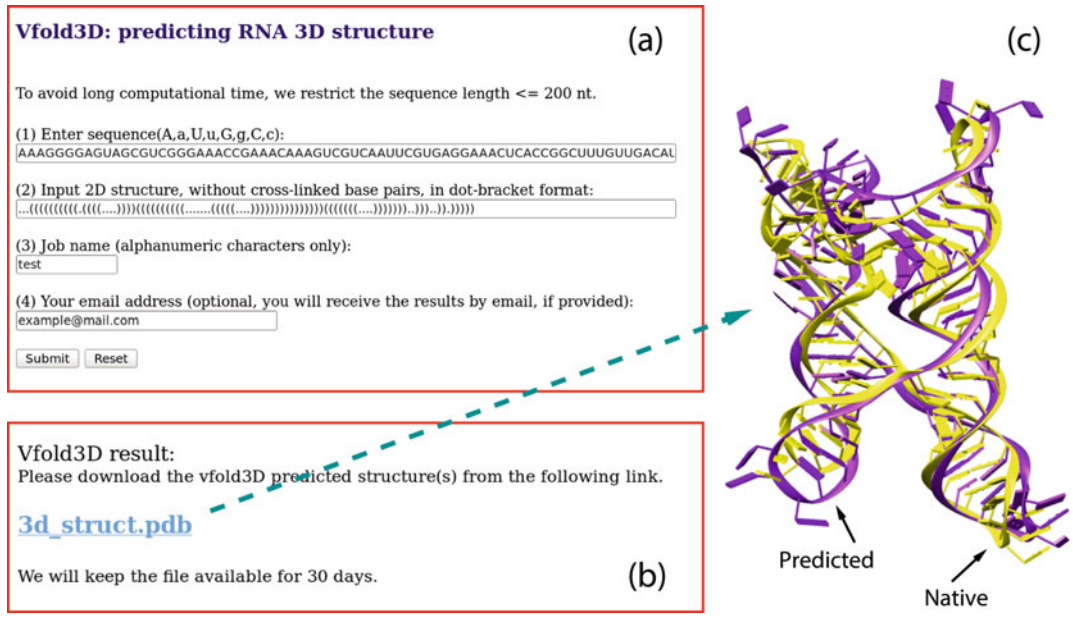


Fig. 4 An example of the Vfold3D prediction. (**a**) Input interface of Vfold3D server. Users have options to input sequence, and 2D structure in dot-bracket format. (**b**) Output interface of Vfold3D server. Users can download predicted 3D structures from the result page. An error message will be displayed if Vfold3D cannot find proper templates for at least one motif. (**c**) The comparison between the predicted and native structures (RMSD = 6.9 A)

5. An error message will be given, if Vfold3D cannot find proper templates for at least one motif.
6. On the result page, Vfold3D outputs the predicted all-atom 3D structure(s) in the PDB format (*see* Fig. 4b).

We use *yybP-ykoY* orphan riboswitch as an example to show how Vfold predicts 3D structures. The Vfold2D predicted 2D structure shows seven helices, one four-way junction, three internal/bulge loops, three hairpin loops, and one 5′-unpaired loop. For the 2D structure from Vfold2D, which contains incorrectly predicted 5 (out of 33) canonical base pairs, the RMSD to the experimentally determined native structure is 6.9 Å (Fig. 4c), which indicates that Vfold3D predicted structure can indeed capture the global fold of the structure, even for 2D structures with minor inaccuracies.

If the (fully correct) native 2D structure (*see* **Note 16**) of the *yybP-ykoY* orphan riboswitch is used as the input for Vfold3D, the RMSD of the predicted 3D structure would be reduced to 3.3 Å. The usage of the A-form helices in Vfold3D, which is slightly different from the helices in the experimentally determined RNA structures, may cause a notable contribution to the RMSD.

4 Notes

1. A survey of the known structures suggests that the virtual bonds (P-C4′ and C4′-P) have bond length of $\sim$ 3.9 Å and bond angle in the range of 90–120°.
2. By enumerating all the possible (sequence-dependent) intra-loop mismatches considered in the Vfold2D algorithm, the Vfold model can partially account for the sequence-dependence of the loop free energy.
3. The usage of all-atom helices can better account for the excluded volume effects between helices and between helices and loops in the loop entropy calculations.
4. The length of each loop is limited to 8 nt for the complete loop virtual-bonded structure ensemble, due to the long computational time for the exhaustive self-avoiding walks.
5. For the RNA 3D structures solved by NMR, only the first model is used to extract motifs for the database.
6. An RMSD cutoff of 1.2 Å (between the standard A-form helix and the helices in real RNA structures) is used for the helix extraction.
7. The fragment assembly-based method only considers the intra-loop structural features, while the motif assembly-based method conserves both the intra- and inter loop interactions within the motifs.

8. Each helix should contain at least 2 base pairs. The helices with single base pair are treated as the intra-motif interactions.
9. Vfold defines the sequence distance $H = \sum_i h^i$ to find the optimal templates. Here, h^i is the hamming distance between nucleotide i in the selected template and the corresponding nucleotide in the target sequence through the following substitution cycles: A → G → C → U, C → U → A → G, G → A → U → C, U → C → G → A.
10. The MD minimization, such as AMBER and NAMD, only causes small RMSD change in 3D structure. Currently, the energy minimization has not been automated in the Vfold3D server.
11. We have pre-tabulated Vfold-derived parameters for the different types of the loops [25, 29–32].
12. There are minor differences between these two sets of energy parameters.
13. The computational time scales with the chain length N as $O(N^6)$ and the memory scales as $O(N^2)$.
14. RNAs often have multiple, heterogeneous conformational distributions with the formation of multiple stable and metastable structures. Therefore, the predicted minimum free energy structures may not always correspond to the native structures, due to the conformational flexibility and the uncertainty in the energy parameters derived by the experiments and theories.
15. The sequence of *yybP-ykoY* orphan riboswitch is: 5′AAAGGGGAGUAGCGUCGGGAAACCGAAACAAAG UCGUCAAUUCGUGAGGAAACUCA CCGGCUUUGUU GACAUACGAAAGUAUGUUUAGCAAGACCU UUCC3′.
16. The native 2D structure: (((((.....((.((((....)))) ((((((((((.....((((((....)))))))))))))))) (((((((....)))))))..)).....)))))..
17. Other RNA secondary structure prediction models, such as Mfold [13], RNAstructure [14], RNAfold [15], and MC-Fold [22], can also be used.

Acknowledgements

This research was supported by NIH grant R01-GM063732.

References

1. Ladd M, Palmer R (1985) Structure determination by X-ray crystallography. Plenum Press, New York, p 71
2. Furtig B, Richter C, Wohnert J, Schwalbe H (2003) NMR spectroscopy of RNA. Chembiochem 4(10):936–962. doi:10.1002/cbic.200300700
3. Bender W, Davidson N (1976) Mapping of poly (A) sequences in the electron microscope reveals unusual structure of type C oncornavirus RNA molecules. Cell 7(4):595–607. doi:10.1016/0092-8674(76)90210-5
4. Proudfoot NJ, Brownlee GG (1976) 3′ noncoding region sequences in eukaryotic messenger RNA. Nat Rev Genet 2(12):211–214. doi:10.1038/263211a0
5. Gardner PP, Giegerich R (2004) A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinform 5:140. doi:10.1186/1471-2105-5-140
6. Mathews DH, Moss WN, Turner DH (2010) Folding and finding RNA secondary structure. Cold Spring Harb Perspect Biol 2:a003665. doi:10.1101/cshperspect.a003665
7. Washietl S (2010) Sequence and structure analysis of noncoding RNAs. Methods Mol Biol 609:285–306. doi:10.1007/978-1-60327-241-4_17
8. Machado-Lima A, del Portillo HA, Durham AM (2008) Computational methods in noncoding RNA research. J Math Biol 56:15–49. doi:10.1007/s00285-007-0122-6
9. Mathews DH, Turner DH (2006) Prediction of RNA secondary structure by free energy minimization. Curr Opin Struct Biol 16:270–278. doi:10.1016/j.sbi.2006.05.010
10. Sato K, Kato Y, Akutsu T, Asai K, Sakakibara Y (2012) DAFS: simultaneous aligning and folding of RNA sequences via dual decomposition. Bioinformatics 28(24):3218–3224. doi:10.1093/bioinformatics/bts612
11. Mathews DH, Turner DH (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. J Mol Biol 317(2):191–203. doi:10.1006/jmbi.2001.5351
12. Tucker BJ, Breaker RR (2005) Riboswitches as versatile gene control elements. Curr Opin Struct Biol 15(3):342–348. doi:10.1016/j.sbi.2005.05.003
13. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406–3415. doi:10.1093/nar/gkg595
14. Bellaousov S, Reuter JS, Seetin MG, Mathews DH (2013) RNAstructure: web servers for RNA secondary structure prediction and analysis. Nucleic Acids Res 41:W471–W474. doi:10.1093/nar/gkt290
15. Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31 (13):3429–3431. doi:10.1093/nar/gkg599
16. Turner DH, Mathews DH (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res 38:D280–D282. doi:10.1093/nar/gkp892
17. Tan RK, Petrov AS, Harvey SC (2006) YUP: a molecular simulation program for coarse-grained and multi-scaled models. J Chem Theory Comput 2:529–540. doi:10.1021/ct050323r
18. Jonikas MA, Radmer RJ, Laederach A, Das R, Pearlman S, Herschlag D, Altman RB (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA 15:189–199. doi:10.1261/rna.1270809
19. Sharma S, Ding F, Dokholyan NV (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. Bioinformatics 24:1951–1952. doi:10.1093/bioinformatics/btn328
20. Xia Z, Bell DR, Shi Y, Ren P (2013) RNA 3D structure prediction by using a coarse-grained model and experimental data. J Phys Chem B 117:3135–3144. doi:10.1021/jp400751w
21. Das R, Karanicolas J, Baker D (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. Nat Methods 7:291–294. doi:10.1038/nmeth.1433
22. Parisien M, Major F (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature 452:51–55. doi:10.1038/nature06684
23. Cao S, Chen S-J (2011) Physics-based de novo prediction of RNA 3D structures. J Phys Chem B 115:4216–4226. doi:10.1021/jp112059y
24. Xu X, Zhao P, Chen S-J (2014) Vfold: a web server for RNA structure and folding thermodynamics prediction. PloS One 9(9):e107504. doi:10.1371/journal.pone.0107504
25. Cao S, Chen S-J (2005) Predicting RNA folding thermodynamics with a reduced chain representation model. RNA **11**:1884–1897. doi:10.1261/rna.2109105
26. Chen S-J (2008) RNA folding: conformational statistics, folding kinetics, and ion

electrostatics. Annu Rev Biophys 37:197–214. doi:10.1146/annurev.biophys.37.032807.125957

27. Ferro DR, Hermans J (1971) A different best rigid-body molecular fit routine. Acta Crystallogr A 33:345–347. doi:10.1107/S0567739477000862
28. Arnott S, Hukins DW, Dover SD (1972) Optimised parameters for RNA double-helices. Biochem Biophys Res Commun 48:1392–1399. doi:10.1016/0006-291X(72)90867-4
29. Cao S, Chen S-J (2006) Predicting RNA psuedoknot folding thermodynamics. Nucleic Acids Res 34:2634–2652. doi:10.1093/nar/gkl346
30. Cao S, Chen S-J (2009) Predicting structures and stabilities for H-type pseudoknots with inter-helix loop. RNA 15:696–706. doi:10.1261/rna.1429009
31. Cao S, Chen S-J (2011) Structure and stability of RNA/RNA kissing complex: with application to HIV dimerization initiation signal. RNA 17:2130–2143. doi:10.1261/rna.026658.111
32. Cao S, Chen S-J (2012) A domain-based model for predicting large and complex pseudoknotted structures. RNA Biol 9:200–211. doi:10.4161/rna.18488
33. Sarver M, Zirbel CL, Stombaugh J, Mokdad A, Leontis NB (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. J Math Biol 56(1–2):215–252. doi:10.1007/s00285-007-0110-x
34. Bindewald E, Hayes R, Yingling YG, Kasprzak W, Shapiro BA (2008) RNAJunction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nanodesign. Nucleic Acids Res 36:D392–D397. doi:10.1093/nar/gkm842
35. Popenda M, Szachniuk M, Blazewicz M, Wasik S, Burke EK, Blazewicz J, Adamiak RW (2010) RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. BMC Bioinformatics 11:231. doi:10.1186/1471-2105-11-231
36. Petrov AI, Zirbel CL, Leontis NB (2013) Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. RNA 19 (10):1327–1340. doi:10.1261/rna.039438.113
37. Darty K, Denise A, Ponty Y (2009) VARNA: interactive drawing and editing of the RNA secondary structure. Bioinformatics 25 (15):1974–1975. doi:10.1093/bioinformatics/btp250
38. Price IR, Gaballa A, Ding F, Helmann JD, Ke A (2015) Mn(2+)-sensing mechanisms of yybP-ykoY orphan riboswitches. Mol Cell 57 (6):1110–1123. doi:10.1016/j.molcel.2015.02.016

Chapter 2

RNA Function Prediction

Yongsheng Li, Juan Xu, Tingting Shao, Yunpeng Zhang, Hong Chen, and Xia Li

Abstract

Recent studies have shown that a considerable proportion of eukaryotic genomes are transcribed as noncoding RNA (ncRNA), and regulatory ncRNAs have attracted much attention from researchers in many fields, especially of microRNA (miRNA) and long noncoding RNA (lncRNA). However, most ncRNAs are functionally uncharacterized due to the difficulty to accurately identify their targets. In this chapter, we first summarize the most recent advances in ncRNA research and their primary function. We then discuss the current state-of-the-art computational methods for predicting RNA functions, which comprise three different categories: miRNA function prediction approaches using target genes, lncRNA function prediction based on the guilt-by-association principle, and RNA function prediction approaches based on competing endogenous RNA partners. We consider that the application of these techniques can provide valuable functional and mechanistic insights into ncRNAs, and that they are crucial steps in future functional studies.

Key words Co-epigenetic modification, Co-expression, Genomic co-location, Competing endogenous RNA, Guilt-by-association principle, lncRNA function prediction, miRNA function prediction, mRNA function prediction, RNA function, Target genes

1 Introduction

The development of high-throughput sequencing technology has made it clear that the transcriptional landscape is far more complex than originally considered. Most genomic sequences can be transcribed into protein-coding RNA (messenger RNA, mRNA) or noncoding RNA (ncRNA). ncRNAs are generally classified into two groups: noncoding housekeeping and regulatory ncRNAs [1]. The latter tend to be expressed at certain stages in an organism's development or during cell differentiation, and they can affect the expression of other genes at the transcriptional or translational levels. Thus, they have been attracted much attention from researchers in many fields. Currently, our understanding of these regulatory ncRNAs is mainly focused on microRNA (miRNA) and long noncoding RNA (lncRNA). miRNAs comprise about 22

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_2,

nucleotides (nt) and they are single-stranded RNAs that have been highly conserved throughout evolution [2]. Since their discovery 20 years ago, miRNAs have attracted much attention in all areas of biological research. In addition, lncRNAs play critical roles in the cell [3]. A lncRNA is a type of RNA comprising more than 200 nt without any apparent protein-coding role [4].

Increasing evidence indicates that most ncRNAs act as regulators that participate in important biological functions in the cell and that they are associated with many types of complex diseases [5]. The application of high-throughput sequencing and recent progress in bioinformatic methods has increased the number of known ncRNAs rapidly. However, the gap between the numbers of identified and functionally characterized molecules is very large. Ultimately, the functionality of an ncRNA should be validated by experimental biological approaches. However, classic methods such as gene knockdown, overexpression, or editing are often not suitable for analyzing an extensive pool of ncRNA candidates. However, it is possible to perform genome-wide investigations and interpret the in silico functionality to narrow the functional search space for many ncRNAs by using computational methods and publicly available datasets. We first give a brief introduction of RNA function in general, especially that of mRNAs, miRNAs, and lncRNAs. Then we describe some commonly used computational approaches for predicting the functions of miRNAs and lncRNAs.

2 RNA Functions

2.1 *mRNAs*

mRNAs are key types of cellular RNAs and they mainly perform functions via their corresponding proteins. Numerous databases describe the well-known functions of the proteins corresponding to mRNAs, such as Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [6] (Fig. 1, top panel). In addition, recent studies have suggested that some mRNAs of protein-coding genes have independent functions, e.g., some mRNAs can act as miRNA sponges to regulate the expression of other mRNAs, which are known as competing endogenous RNAs (ceRNAs).

2.2 *miRNAs*

miRNAs are highly important regulatory molecules at the posttranscriptional level. miRNAs are transcribed by RNA polymerase II as primary miRNAs (pri-miRNAs), which are then processed by Drosha to produce thermodynamically stable hairpin structures known as pre-miRNAs. These pre-miRNAs are then exported into the cytoplasm by Exportin-5 and processed further by the RNAase III enzyme Dicer to form miRNA duplexes. miRNAs can negatively regulate gene expression via partial base pairing with target mRNAs to influence the mRNA degradation process or

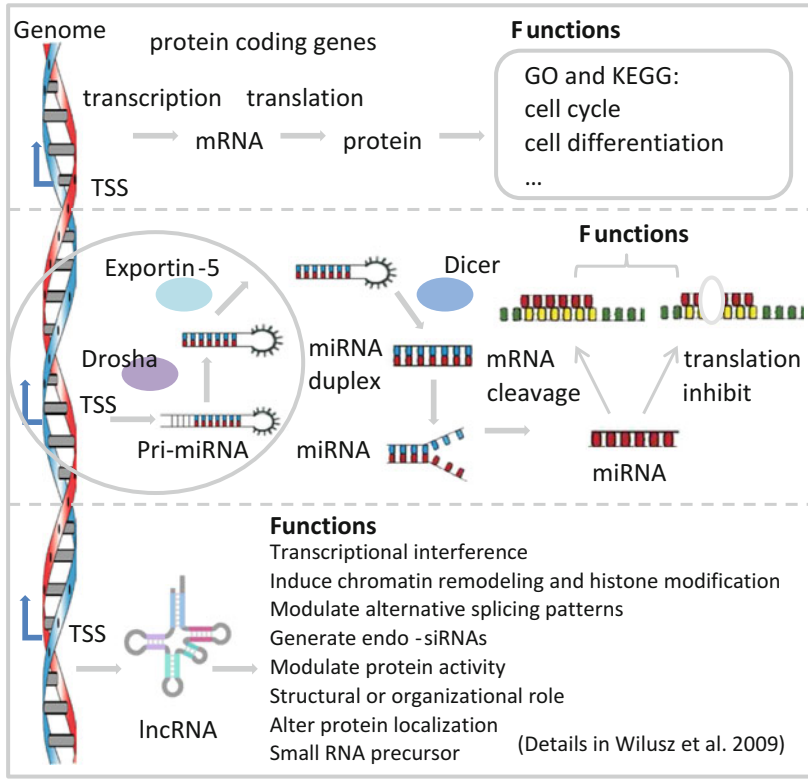


Fig. 1 RNAs perform functions via different mechanisms. A genomic sequence can be transcribed into protein-coding RNA (mRNA), miRNA, or lncRNA

repress translation (Fig. 1, middle panel) [7]. There is evidence that about 60% of human protein-coding genes can be targeted by miRNAs and they can perform multiple functions via their target genes [2].

2.3 lncRNAs

Most of the functions of lncRNAs are still unknown, but it has been shown that lncRNAs can regulate gene expression at the epigenetic, transcriptional, and posttranscriptional levels, e.g., by genetic imprinting, chromatin remodeling, coregulation of transcription factors, splicing regulation, mRNA decay, and translational regulation (Fig. 1, bottom panel) [4]. They also harbor miRNA response elements (MREs) and can act as sponges for miRNAs.

3 miRNA Function Prediction Approaches Using Target Genes

3.1 General Function Prediction

The most commonly used computational method for identifying the functions of miRNAs is based on their target genes. Thus, one of the biggest challenges is identifying the targets regulated by miRNAs. During the last few years, various computational

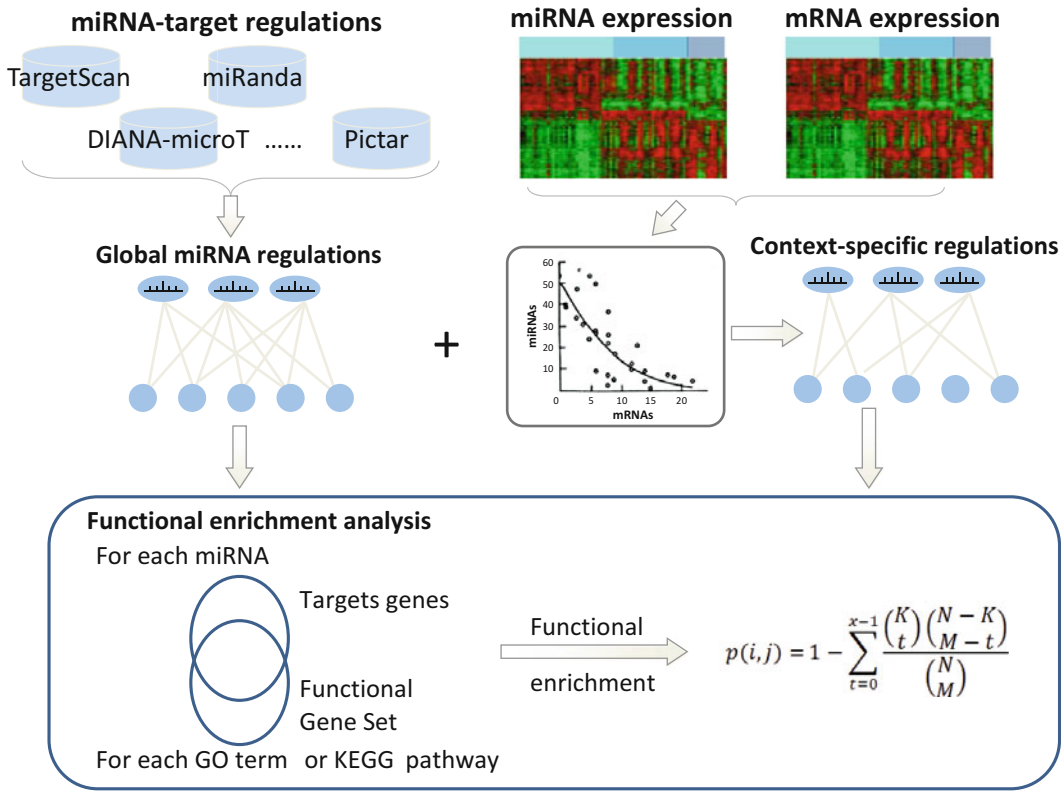


Fig. 2 miRNA function prediction approaches using target genes. miRNA functional prediction approaches using target genes can be divided into two classes: general function prediction and context-specific function prediction

approaches have been designed to predict the target genes of miRNA sequences mainly by considering the degree of sequence complementarity, such as TargetScan and miRanda [8–11] (Fig. 2). However, miRNA target identification is challenging due to the imperfect nature of base pairing and because the rules of targeting are not completely understood. On the other hand, significant efforts have been made to determine biologically relevant miRNA–target interactions using high-throughput experimental approaches. In particular, the use of crosslinking and Argonaute immunoprecipitation coupled with high-throughput sequencing can identify endogenous genome-wide interaction maps for miRNAs [12], thereby providing an alternative to sequencing and in silico prediction strategies. The combination of computational with experimental approaches can refine the computational predictions.

Next, we consider how to use these target genes to predict the functions of miRNAs. Functional enrichment analysis is generally used to link ncRNAs and their related functions. For example, for an miRNA of interest i, we can determine the corresponding target genes $T_i = \{g_1, g_2, g_3, \ldots, g_M\}$. In addition, the genes with a

specific function can be obtained from the GO or KEGG pathway, which are denoted as $F_j = \{g_1, g_2, g_3, \ldots, g_K\}$. We can then use statistical tests to compute the significance of function enrichment, e.g., using the hypergeometric test, as follows:

$$p(i,j) = 1 - F(x|N,K,M) = 1 - \sum_{t=0}^{x-1} \frac{\binom{K}{t}\binom{N-K}{M-t}}{\binom{N}{M}}$$

where N is the number of all genes, K is the number of genes annotated in the specific GO term or KEGG pathway j, M is the size of the target gene for miRNA i, and x is the number of targets annotated to a GO term or KEGG pathway j. We can compute the p-value for each GO term or KEGG pathway, and these p-values may be corrected by multiple testing adjustments. At a given significance level, we can obtain the functional terms or KEGG pathways for each specific miRNA.

3.2 Context-Specific Function Prediction

The global methods for function analysis mentioned above are focused on miRNA–mRNA regulation at a global level. However, a major limitation of these approaches is that we might expect miRNA regulation to be reprogrammed in different biological contexts. To address this limitation, computational methods have been proposed for modeling context-specific miRNA–mRNA regulation and using these specific targets for predicting miRNA functions [13, 14]. This process requires paired miRNA and mRNA expression profiles for the same samples. In particular, to identify functional regulation from miRNA to mRNA, we can combine the computational target predictions at the sequence level and the inverse expression relationships between miRNA and mRNA expression in a specific context (Fig. 2). miRNAs tend to downregulate their target mRNAs, so the expression profiles of genuinely interacting pairs are expected to be anti-correlated. Thus, the correlation coefficients or a linear regression model can be used to estimate the correlation between the expression of each miRNA and all of the protein-coding genes. After selecting cutoffs for the correlation coefficient and p-value, the context-specific target genes can be identified for each miRNA. We can then identify the function of a miRNA based on the functional enrichment analysis described above.

4 lncRNA Function Prediction Based on the Guilt-by-Association Principle

At present, only a limited number of lncRNAs have been characterized in detail. The guilt-by-association principle has been used widely to infer the functions of lncRNAs mainly via their co-expression, co-location, or co-epigenetic regulation with protein-coding genes.

4.1 Computational Annotation of lncRNA Functions Based on lncRNA-Gene Co-expression and Genomic Co-location

Similar to miRNA functional prediction, it is essential to identify the target genes of lncRNAs. Gene expression information is used mainly to detect potential regulatory targets. In general, a protein-coding gene is considered to be a target of an lncRNA if it is differentially expressed after knocking down or overexpressing the lncRNA (Fig. 3b), and these differentially expressed genes can be used to perform functional enrichment analysis, as described above for miRNAs [15]. However, the number of lncRNA knockdown/overexpression experiments is still limited, thereby hindering the identification of regulated genes to increase the number of characterized lncRNAs. Alternatively, the most intuitive and commonly used method is to identify genes with correlated expression. Predicting lncRNA functions based on lncRNA-mRNA co-expression assumes that if an lncRNA regulates a gene, then the expression of the lncRNA is significantly correlated with the expression of the mRNA, where this type of method requires paired expression profiles for lncRNAs and mRNAs as input data. The RNA-seq technique can quantify the transcribed molecules in various samples, thereby providing an ideal data source for determining genome-wide lncRNA and mRNA expression profiles. Another source

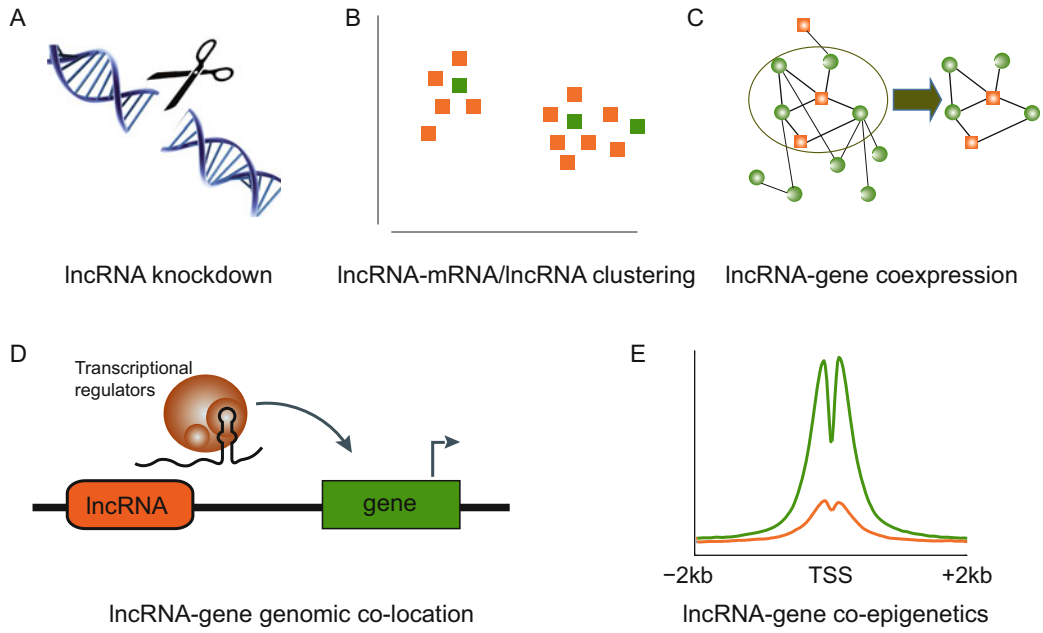


Fig. 3 lncRNA function prediction based on the guilt-by-association principle. Inferring lncRNA functions via lncRNA knockdown or overexpression experiments (**a**). Characterizing lncRNAs based on an lncRNA with a known function or protein-coding genes in the same cluster (**b**). Predicting the functions of lncRNAs based on an lncRNA-gene co-expression network or network module (**c**). Computational annotation of lncRNA functions based on lncRNA-gene genomic co-location (**d**). lncRNAs share common chromatin patterns with protein-coding genes around their transcription start sites (+2 kb) (**e**). Green represents the protein-coding gene and orange represents the lncRNA, respectively

comprises the microarray data obtained from microarray re-annotation or lncRNA arrays, where the probes are specifically designed for lncRNAs [16]. The paired expression profiles of lncRNAs and mRNAs can be constructed after obtaining these expression data. Genome-wide clustering analysis of lncRNA and mRNA expression profiles may group uncharacterized lncRNAs, where the lncRNAs with known functions and protein-coding genes may be assigned to various different clusters, whereas genes with similar expression profile are clustered into one gene group. The functions of uncharacterized lncRNAs can be predicted based on lncRNAs with known functions or protein-coding genes in the same cluster because they are more likely to be involved in the same biological process (Fig. 3b). A linear regression model or other methods can be used to identify protein-coding genes that are co-expressed with lncRNAs. The following two strategies are employed to functionally annotate lncRNAs. The first strategy is based on functional enrichment analysis. Thus, for each lncRNA, the co-expressed protein-coding genes are subjected to GO and KEGG function enrichment analyses as miRNAs and the enriched functions are also annotated for the lncRNA (Fig. 3c) [17]. The second approach is based on a network model and it uses specific algorithms. A co-expression network of lncRNAs and mRNAs can be constructed by assembling all the significantly co-expressed lncRNA-mRNA pairs [18]. Computational methods have been developed to predict the candidate functions of lncRNAs in the network model. For example, different network modules may be detected in a given co-expression network and the genes in the same module are considered to participate in the same biological function. The aim of these methods is to detect co-expressed modules.

In terms of the distance of regulation, lncRNAs can regulate transcription in *cis* and *trans*, where *cis*-acting lncRNAs control the expression of protein-coding genes located in the vicinity of their transcriptional start sites [19]. lncRNAs can regulate transcription in *cis* by recruiting specific transcriptional regulators to nearby protein-coding genes. Thus, it is possible to infer the functions of lncRNAs by identifying co-located lncRNA-coding gene pairs (Fig. 3d) [20]. Indeed, it has been demonstrated that if an lncRNA is co-expressed with a nearby coding gene, then the two genes are frequently separated by a distance of less than 10 kb in the linear genome.

4.2 Computational Annotation of lncRNA Functions Based on lncRNA-Gene Co-epigenetic Modifications

It has been established that the transcription of lncRNAs is also tightly regulated by epigenetic modification in a similar manner to that of protein-coding genes [21]. Moreover, groups of functionally related genes can be further distinguished at the chromatin level, although they have similar expression patterns [22]. Thus, epigenetic modifications can be used to divide co-expressed gene sets into subgroups that tend to be involved with the same biological processes. Indeed, Li et al. found that the co-expression of protein-coding genes only provides a relatively narrow range for

function prediction [23], but surprisingly, the epigenetic modifications have much greater similarity. Moreover, the vast majority of GO terms with significantly high co-expression also share high chromatin similarity. Thus, it is reasonable to predict lncRNA functions by shifting from co-expression to co-epigenetic modification, or by integrating both. An integrative model was proposed to predict the functions of lncRNAs by combining the chromatin state with expression patterns (Fig. 3e). Thus, by exploiting the wealth of datasets obtained from the ENCODE project, the genome-wide expression profiles and nine chromatin profiles of lncRNAs and genes were compiled. For each GO term, the nearest shrunken centroid algorithm was used to construct a classifier to distinguish genes annotated with the function from randomly selected gene sets. Each feature profile was considered and thus the chromatin features could capture the functions of genes with high power. Using the trained model, the probable functions of more than 97% of human lncRNAs were predicted.

5 RNA Function Prediction Approaches Based on ceRNA Partners

As described above, both mRNAs and lncRNAs can talk to each other using their MREs, thereby acting as ceRNAs. Thus, the ceRNA activity forms a large-scale regulatory network across the transcriptome to greatly expand the functional genetic information in the human genome. The ceRNA partners of target RNAs can be used to predict the functions of RNAs in a similar manner to the methods used for predicting the functions of miRNA based on their target genes. Here we describe the current state of the art in terms of RNA functional prediction approaches based on ceRNA partners, where we can divide the ceRNA recognition methods into three classes: sharing miRNA-based approaches, sharing miRNA and co-expression-based prediction methods, and dysregulated ceRNA–ceRNA interaction-based approaches (Fig. 4).

5.1 Sharing miRNA-Based Prediction Methods

RNAs that share MREs compete for miRNA binding to regulate each other. Sharing miRNAs is the most typical feature of ceRNA pairs and possibly the most widely used computational prediction method (Fig. 4a). The simplest way of predicting ceRNA partners for specific RNAs targeted by miRNAs is to examine the degree of MRE co-occurrence in the mRNAs on a genome-wide scale [24]. It has been shown that trans-regulatory ceRNA crosstalk increases with the number of miRNAs shared by RNAs [25]. Thus, the number of miRNAs shared between RNAs with statistical significance must be considered. The hypergeometric test is also used to measure whether two RNA components share significant miRNAs. starBase v2.0 developed by Li et al. uses the hypergeometric test to predict ceRNA pairs among mRNAs, lncRNAs, circRNAs, and pseudogenes based on the idea of sharing miRNAs. They also

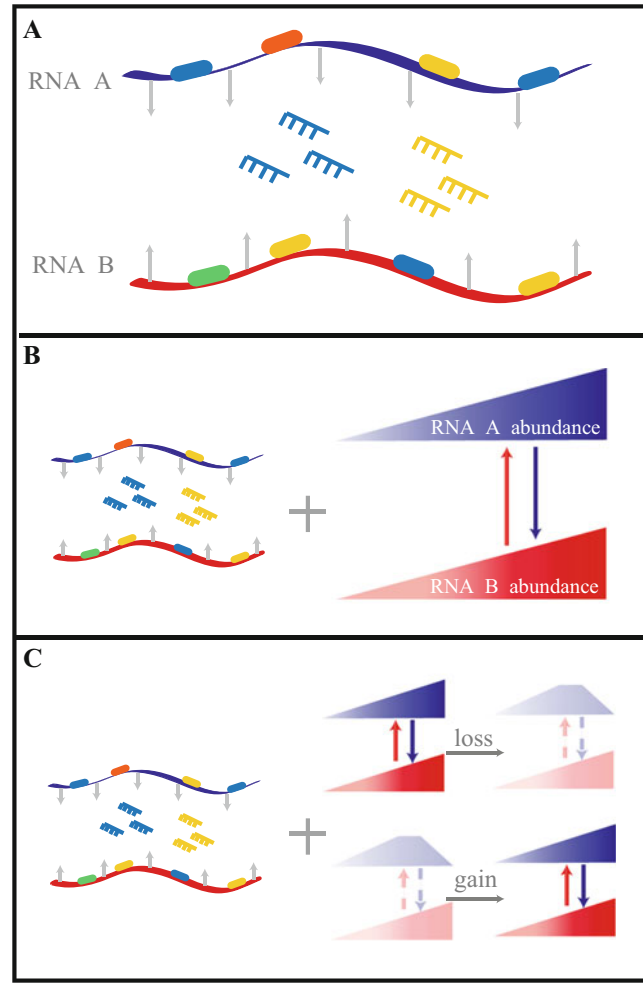


Fig. 4 RNA function prediction approaches based on ceRNA partners. RNA A (*blue*) and B (*red*) are a pair of ceRNAs that share MREs (*boxes*) for two miRNAs (*blue* and *yellow*) (**a**). Sharing miRNAs is the most typical feature of ceRNA pairs and possibly the most widely used computational prediction method. A change in the abundance of one ceRNA will have a similar effect on the level of the other ceRNA (**b**). Some algorithms have been developed based on both sharing miRNAs and co-expression. The ceRNA partners are identified based on sharing miRNAs where a change in the correlated expression of the ceRNA pair in cancerous tissue is compared with that in normal tissue to calculate the extent of dysregulation (gain and loss ceRNAs) (**c**)

developed ceRNA function web servers to predict the functions of lncRNAs based on their corresponding ceRNA partners [26].

5.2 Sharing miRNA and Co-expression-Based Prediction Methods

The RNAs under consideration co-regulate each other's expression level in the simplest scenario of crosstalk between two ceRNAs competing for shared miRNAs. In the steady state, the ceRNAs and targeted miRNAs are in equilibrium. After the abundance of

one ceRNA changes, there will be a similar effect on the level of the other ceRNA, i.e., in one ceRNA pair, overexpression of one of the target RNAs would reduce the concentration of free miRNAs, thereby increasing the expression of the other target RNA. By contrast, decreasing the expression of the target RNA would increase the miRNAs available to bind with the other target RNA, thereby suppressing its expression level [27]. Thus, the expression levels of ceRNA pairs are positively correlated with each other. There is evidence for a strong positive correlation between the expression levels of individual ceRNA components [28]. Algorithms have been developed based on both sharing miRNAs and co-expression (Fig. 4b). For example, Xu et al. developed a two-step method for predicting the ceRNA–ceRNA interaction landscape across 20 cancer types [29]. First, their method computes the significance of shared miRNAs for each possible RNA pair using a hypergeometric test and the number of shared miRNAs is required to at least three. Second, the Pearson's correlation coefficient for each candidate ceRNA pair is computed to further screen for positively co-expressed RNAs. Recently, this method was improved by adding the similarity of the miRNA regulation strength [30]. The ceRNAs of lncRNAs in tissue developmental processes in rhesus apes were identified using the new algorithm and genome-wide predictions of the functions of lncRNAs were obtained based on the ceRNA partners of lncRNAs. For example, lncRNA XLOC_062139 was predicted to interact with ten ceRNA partners that are known brain development-related genes, and thus the lncRNA was inferred as a potential regulator involved in brain development. In addition, mutual information, conditional mutual information, and partial correlation coefficients have been used to predict the ceRNAs of lncRNAs [31].

5.3 Dysregulated ceRNA–ceRNA Interaction-Based Prediction Methods

The functions that RNAs play with their ceRNA partners can change in different conditions. Normal ceRNA regulation is needed for the correct functioning of a cell. Thus, disrupting these ceRNA pairs may promote the development of disease. Detecting this disruption can help us to understand the functions of RNAs in a specific background better than the commonly used ceRNA analysis that only considers one condition. Shao et al. developed a computational approach for identifying dysregulated ceRNA–ceRNA interactions by integrating miRNA regulation with RNA-seq data from cancerous and normal tissues. They identified the ceRNA partners based on sharing miRNAs and the change in the correlated expression of the ceRNA pair in cancerous tissue compared with normal tissue was calculated to determine the extent of dysregulation (Fig. 4c). Some lncRNAs and pseudogenes with potentially dysregulated ceRNA partners affect the levels of competing RNAs that underpin cancer development. For example, most of the dysregulated ceRNA partners of the HSPD1-2P RNA

are involved in cell cycle control and they are over-expressed in lung adenocarcinoma. This RNA may play a role in making the cell cycle more active to promote the development of lung adenocarcinoma via these dysregulated ceRNA interactions [32].

6 Conclusions and Future Prospects

lncRNAs and miRNAs are being discovered continually due to the increasing application of high-throughput RNA sequencing methods. However, the gap between the number of ncRNA identified and functionally characterized ncRNAs remains very large. In this chapter, we discussed the functional modes of three different types of RNAs and described the typical computational methods used for predicting the functions of miRNAs and lncRNAs. We also described ceRNA identification methods that reflect their complex regulation. These methods have been widely accepted and used, but assessments of these programs are still needed. It has been shown that composite methods based on diverse features for assessing different functional aspects are most likely to succeed.

Like transcription factors, miRNAs and lncRNAs are abundant classes of gene regulatory molecules in animal cell. Systematically applying computational prediction methods is a fundamental step in the functional characterization of ncRNAs. It is possible to investigate and interpret the functions of ncRNAs in silico to narrow the search space for functional experimental validations of individual or groups of ncRNAs by using publicly available datasets and computational prediction methods. We consider that the application of these techniques can provide valuable functional and mechanistic insights into ncRNAs, and thus they are essential steps in subsequent functional studies.

References

1. Tano K, Akimitsu N (2012) Long non-coding RNAs in cancer progression. Front Genet 3:219
2. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116 (2):281–297
3. Li J, Xuan Z, Liu C (2013) Long non-coding RNAs and complex human diseases. Int J Mol Sci 14(9):18790–18808
4. Wilusz JE, Sunwoo H, Spector DL (2009) Long noncoding RNAs: functional surprises from the RNA world. Genes Dev 23 (13):1494–1504
5. Esteller M (2011) Non-coding RNAs in human disease. Nat Rev Genet 12 (12):861–874
6. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al (2000) Gene ontology: tool for the unification of biology. The Gene ontology consortium. Nat Genet 25 (1):25–29
7. Ambros V (2004) The functions of animal microRNAs. Nature 431(7006):350–355
8. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R (2004) Fast and effective prediction of microRNA/target duplexes. RNA 10 (10):1507–1517
9. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. Nat Genet 39 (10):1278–1284

10. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS (2004) Human MicroRNA targets. PLoS Biol 2(11):e363
11. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB (2003) Prediction of mammalian microRNA targets. Cell 115(7):787–798
12. Clark PM, Loher P, Quann K, Brody J, Londin ER, Rigoutsos I (2014) Argonaute CLIP-Seq reveals miRNA targetome diversity across tissue types. Sci Rep 4:5947
13. Li Y, Xu J, Chen H, Bai J, Li S, Zhao Z, Shao T, Jiang T, Ren H, Kang C et al (2013) Comprehensive analysis of the functional microRNA-mRNA regulatory network identifies miRNA signatures associated with glioma malignant progression. Nucleic Acids Res 41(22):e203
14. Xu J, Li CX, Lv JY, Li YS, Xiao Y, Shao TT, Huo X, Li X, Zou Y, Han QL et al (2011) Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer. Mol Cancer Ther 10(10):1857–1866
15. Jiang Q, Wang J, Wu X, Ma R, Zhang T, Jin S, Han Z, Tan R, Peng J, Liu G et al (2015) LncRNA2Target: a database for differentially expressed genes after lncRNA knockdown or overexpression. Nucleic Acids Res 43(Database issue):D193–D196
16. Du Z, Fei T, Verhaak RG, Su Z, Zhang Y, Brown M, Chen Y, Liu XS (2013) Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. Nat Struct Mol Biol 20(7):908–913
17. Zhao Z, Bai J, Wu A, Wang Y, Zhang J, Wang Z, Li Y, Xu J, Li X (2015) Co-LncRNA: investigating the lncRNA combinatorial effects in GO annotations and KEGG pathways based on human RNA-Seq data. Database (Oxford) 2015
18. Guo X, Gao L, Liao Q, Xiao H, Ma X, Yang X, Luo H, Zhao G, Bu D, Jiao F et al (2013) Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. Nucleic Acids Res 41(2):e35
19. Fatica A, Bozzoni I (2014) Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet 15(1):7–21
20. Liao Q, Liu C, Yuan X, Kang S, Miao R, Xiao H, Zhao G, Luo H, Bu D, Zhao H et al (2011) Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucleic Acids Res 39 (9):3864–3878
21. Sati S, Ghosh S, Jain V, Scaria V, Sengupta S (2012) Genome-wide analysis reveals distinct patterns of epigenetic features in long non-coding RNA loci. Nucleic Acids Res 40 (20):10018–10031
22. Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, Ding H, Wylie JN, Pico AR, Capra JA et al (2012) Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. Cell 151(1):206–220
23. Li Y, Chen H, Pan T, Jiang C, Zhao Z, Wang Z, Zhang J, Xu J, Li X (2015) LncRNA ontology: inferring lncRNA functions based on chromatin states and expression patterns. Oncotarget 6(37):39793–39805
24. Sarver AL, Subramanian S (2012) Competing endogenous RNA database. Bioinformation 8 (15):731–733
25. Tay Y, Kats L, Salmena L, Weiss D, Tan SM, Ala U, Karreth F, Poliseno L, Provero P, Di Cunto F et al (2011) Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. Cell 147(2):344–357
26. Li JH, Liu S, Zhou H, Qu LH, Yang JH (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res 42(Database issue):D92–D97
27. Marques AC, Tan J, Ponting CP (2011) Wrangling for microRNAs provokes much crosstalk. Genome Biol 12(11):132
28. Sumazin P, Yang X, Chiu HS, Chung WJ, Iyer A, Llobet-Navas D, Rajbhandari P, Bansal M, Guarnieri P, Silva J et al (2011) An extensive microRNA-mediated network of RNA–RNA interactions regulates established oncogenic pathways in glioblastoma. Cell 147(2):370–381
29. Xu J, Li Y, Lu J, Pan T, Ding N, Wang Z, Shao T, Zhang J, Wang L, Li X (2015) The mRNA related ceRNA-ceRNA landscape and significance across 20 major cancer types. Nucleic Acids Res 43(17):8169–8182
30. Xu J, Feng L, Han Z, Li Y, Wu A, Shao T, Ding N, Li L, Deng W, Di X et al (2016) Extensive ceRNA–ceRNA interaction networks mediated by miRNAs regulate development in multiple rhesus tissues. Nucleic Acids Res 44:9438–9451
31. Paci P, Colombo T, Farina L (2014) Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer. BMC Syst Biol 8:83
32. Shao T, Wu A, Chen J, Chen H, Lu J, Bai J, Li Y, Xu J, Li X (2015) Identification of module biomarkers from the dysregulated ceRNA–-ceRNA interaction network in lung adenocarcinoma. Mol BioSyst 11(11):3048–3058

Chapter 3

Computational Prediction of Novel miRNAs from Genome-Wide Data

Georgina Stegmayer, Cristian Yones, Laura Kamenetzky, Natalia Macchiaroli, and Diego H. Milone

Abstract

The computational prediction of novel microRNAs (miRNAs) within a full genome involves identifying sequences having the highest chance of being bona fide miRNA precursors (pre-miRNAs). These sequences are usually named candidates to miRNA. The well-known pre-miRNAs are usually only a few in comparison to the hundreds of thousands of potential candidates to miRNA that have to be analyzed. Although the selection of positive labeled examples is straightforward, it is very difficult to build a set of negative examples in order to obtain a good set of training samples for a supervised method. In this chapter we describe an approach to this problem, based on the unsupervised clustering of unlabeled sequences from genome-wide data, and the well-known miRNA precursors for the organism under study. Therefore, the protocol developed allows for quick identification of the best candidates to miRNA as those sequences clustered together with known precursors.

Key words MicroRNAs prediction, Genome-wide data, Unsupervised model, Clustering, Self-organizing map, High class imbalance

1 Introduction

MicroRNAs (miRNAs) are a class of small noncoding RNA molecules, present in both animals and plants, with a major role in regulation of gene expression [1]. Many studies have shown that miRNAs are implied in several important processes, for example, in cancer progression [2] as well as in viral infection progress [3] and parasites development [4]. Given their role in promoting or inhibiting certain diseases and infections, the discovery of new miRNAs is of high interest today. MiRNA precursors (pre-miRNAs, also known as hairpins) generated during biogenesis have well-known RNA secondary structures that have allowed the development of computational algorithms for their identification. They typically exhibit a stem-loop structure or hairpin, with few internal loops or asymmetric bulges. Since large amount of similar hairpins can be

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_3,

folded in a given genome, the identification of those structures having the highest chance of being bona fide pre-miRNAs should be addressed. Due to the difficulty in systematically detecting pre-miRNAs by existing experimental techniques, which have proven to be time consuming and costly, computational methods play an important role nowadays in the identification of novel miRNAs [5, 6]. Machine learning methods essentially identify hairpin structures in noncoding and non-repetitive regions of the genome that are characteristics of miRNA precursor sequences, using structures, properties, and features of well-known pre-miRNAs during the learning processes to discriminate between true predictions and false positives [7].

In a realistic scenario, when genome-wide data is used, a huge imbalance is often present between the positive class (a few known pre-miRNAs) and the unlabeled data (hundreds of thousands sequences). This important fact may lead to overlearning the majority class and/or incorrect assessment of classification performance. This means that most existing supervised proposals, although reporting very high accuracies, cannot be really trusted in practical situations.

In this chapter we present a protocol to predict novel pre-miRNAs from genome-wide data, with a classifier based on unsupervised learning. The model can predict the best candidates to pre-miRNAs, as sequences are clustered together with the well-known pre-miRNAs of the genomics data under study. This way, the very-hard to build negative artificial examples must not be defined, making it useful to work with genome-wide data from any organism.

2 Materials

2.1 Input Data

- genomic DNA: A fasta file of genomic DNA (for example, genome.fa), with an entry for each chromosome. The genomics data will be mined to identify the best miRNA precursors.
- pre-miRNAs: A fasta file of known pre-miRNA sequences. These sequences are retrieved from specialized databases or reported in the literature as experimentally validated. These pre-miRNAs could be from the organism under study or a phylogenetically related one.
- other known non-miRNA RNA sequences (optional): A fasta file of CDSs, tRNAs, rRNAs, non-coding RNAs, and other non-miRNA sequences. These sequences can be used to filter out known other non-miRNA RNAs.

2.2 Software

- Einverted (EMBOSS package). Program for finding inverted repeats in nucleotide sequences and genome folding. Available free from emboss.sourceforge.net/download/.

- RNA fold. This program reads RNA sequences, calculates their minimum free energy (MFE) structure, and prints the MFE structure in bracket notation and its free energy. It can be downloaded from www.tbi.univie.ac.at/RNA/RNAfold.1.html.
- MiRcheck. Scripts to call and process einverted and RNAfold outputs. Available free from bartellab.wi.mit.edu/software.html.
- BLAST. This program finds regions of similarity between biological sequences. Available at ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/.
- miRNA-SOM. This is a tool for the discovery of pre-miRNAs from genome-wide data. Available at sourceforge.net/projects/sourcesinc/files/mirnasom/ (download version 23).
- miRNAfe (optional). It is a comprehensive tool to extract features from RNA sequences, providing almost all state-of-the-art feature extraction methods used today in several works from different authors. Available at fich.unl.edu.ar/sinc/blog/web-demo/mirnafe/.

3 Methods

This section shows in detail the individual steps necessary to carry out the pipeline proposed for the analysis of raw genome-wide data, which is presented in Fig. 1. Each step of the pipeline will be described and exemplified with linux commands.[1] Before beginning, the following software must be installed:

- Install einverted:

```
sudo apt-get install emboss
```

- Install RNAfold:

```
sudo apt-get install vienna-rna
```

3.1 Cut and Fold Genome-Wide Data

The input genome-wide data (a multi-fasta file named, for example, genome.fa) is pre-processed by miRcheck scripts, which calls einverted and RNAfold [8]. These steps can be done as follows:

- Cut full genome into sequences: the original `run_einverted.pl` script from miRcheck can be used, but previously the gap penalty and other thresholds of einverted must be configured (*see* **Notes 1** and **2**). A modified version of the script with these parameters is provided in the utils folder of miRNA-SOM (version 23). With the modified script, the following linux command can be used to run einverted:

```
./run_einverted.pl genome.fa genIR
```

[1] Command-line examples for Ubuntu Linux.

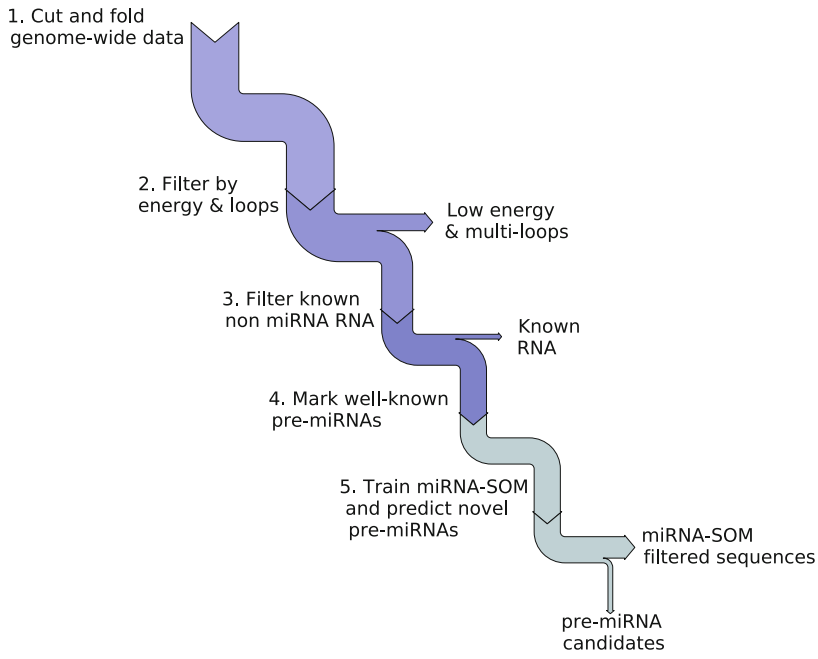


Fig. 1 Flow of the pipeline for novel pre-miRNA discovery from genome-wide data

- If the fasta file has extra information apart from the chromosomes (for example, mitochondrial DNA), it should be disregarded, leaving the chromosomes information only. For example, suppose that genome.fa has a particular string (such as `Chr_<number>`) that identifies chromosomes. Then, you can run:

```
cat genIR | grep Chr_ > genIR_chr
```

- Folding sequences: this step can be done by editing `fold_inverted_repeats.pl`, adding RNAfold options to produce structures without lonely pairs (noLP) and avoid the generation of postscript drawings (noPS).[2] After that, you can run:

```
./fold_inverted_repeats.pl genIR_chr genome.fa genIR_chr_f
```

3.2 Filter by Energy and Loops

The sequences obtained in the previous step, from the raw genome-wide data cut and folding procedure, must be filtered to improve prediction. Two filters can be applied: a minimum free energy (MFE) threshold of −20 according to the miRNA biogenesis model [1], and multi-loops sequences can be discarded, obtaining a reduced fasta file. This step can be done by running the script:

```
filterle.m
```

[2] A modified version of the script is also provided in the utils folder of miRNA-SOM version 23.

provided with the source code of miRNA-SOM. Inside this matlab script, the mentioned filters are applied and a fasta file named `all_folded_selected_le.fa` is obtained as output.

In this step, the script `filterle.m` also extracts the following features, that are the most commonly used in literature for pre-miRNA prediction [6]:

- Triplets [9]: combines the local contiguous structures with sequence information to characterize the hairpin structure. This feature focuses on the information of every 3 adjacent nucleotides. In the predicted secondary structure, there are only two status for each nucleotide, paired or unpaired, indicated by brackets, "(" or ")", and dots, respectively. The left bracket "(" means that the paired nucleotide is located near the 5′-end and can be paired with another nucleotide at the 3′-end, which is indicated by a right bracket ")". For any 3 adjacent nucleotides, there are 2^3 possible structure compositions: `(((`, `((.`, `(..`, `(.(`, `.((`, `.(.`, `..(` and `...`. Considering the middle nucleotide among the 3, there are 32 possible structure-sequence combinations, which are denoted as `U(((`, `A((.`, etc.
- MFE value [10]: minimum free energy when folding; and
- Sequence length: count of the length of the nucleic acid string.

All these features are saved in the data folder of miRNA-SOM, to train the model as detailed in 3.5. Additionally, any number of features can be extracted and used to train the miRNA-SOM classifier. The web-tool miRNAfe [11], which is a comprehensive tool to extract features from RNA sequences, can be used for features extraction. It provides almost all state-of-the-art feature extraction methods used today in several works from different authors.

3.3 Filter Known non-miRNA RNA

This is an optional step. If a fasta file of CDS, tRNAs, rRNAs and long non-coding RNAs, as well as any other non-miRNA sequences of the organism under study is available (for example, in a file named `known_rna.fa`), they can be used to filter out known non-miRNAs. This can be done by using BLAST [12] with the following linux script:

```
./delkrna.sh known_rna.fa all_folded_selected_le.fa all_folded_to_remove.csv
```

This script is also provided with miRNA-SOM. It generates the file *all_folded_to_remove.csv*, which indicates the indexes of the sequences that must be removed from *all_folded_selected_le.fa*.

3.4 Mark well-known pre-miRNAs

As a result of the previous steps, the files `all_folded_selected_le.fa` and `all_folded_to_remove.csv` are obtained. The first one includes sequences that correspond to well-known pre-miRNAs of the organism under study. These known pre-miRNAs can be identified after a BLAST match against the microRNA hairpins deposited in the most recent version of miRBase,[3] and put together into a multi-fasta file, for example named `mirnas.fa`.

These sequences must be labeled as positive class in order to properly train the miRNA-SOM classifier. This step can be done this way:

```
./selmirs.sh mirnas.fa all_folded_selected_le.fa all_folded_known_mirna.csv
```

This script is also provided with miRNA-SOM. It generates the file `all_folded_known_mirna.csv`, which has the indexes of the sequences that correspond to well-known pre-miRNAs in `all_folded_selected_le.fa`.

3.5 Train miRNA-SOM and Predict Novel pre-miRNAs

The `mainsom.m` script provided in miRNA-SOM trains the SOM classifier [13] (shown in Fig. 2). It learns the labeled sequences as positive class, and identifies novel candidates to pre-miRNAs. When this main script is run, the miRNA-SOM classifier is trained according to the Algorithm shown in Fig. 3, where the following notation is used: G_ℓ and G_u are the labeled and unlabeled input training sequences, respectively, extracted from the input genome-wide data and represented by a feature vector (steps 1–4 of the pipeline of Fig. 1). Labeled input sequences correspond to well-known pre-miRNAs; n is the initial map size ($n \times n$ neurons); and $h_{\max}$ is the maximum deep level.

The miRNA-SOM model training and prediction involves the following steps. While the maximum deep level of SOMs has not been reached (line 4), a SOM map is trained at each level (line 5). The top level SOM, at $h = 1$, is set to the initial map size (*see* **Note 3**) and trained with all input training data (labeled and unlabeled data). During training, each input data point is assigned to a map unit (neuron) according to the minimum Euclidean distance between the feature vector representing each sequence and each neuron centroid. Neurons are labeled by taking into account the labeled data only, as follows: if there is at least one labeled input sequence in a neuron (line 6), then this neuron is labeled as a miRNA-neuron, no matter how many other unlabeled data points are clustered there as well. Then, only sequences clustered on miRNA-neurons pass to the next level (line 8). After training all SOM levels, up to $h_{\max}$, only the sequences that are clustered into labeled miRNA-neurons at the deepest level ($h_{\max}$)

[3] http://www.mirbase.org/.

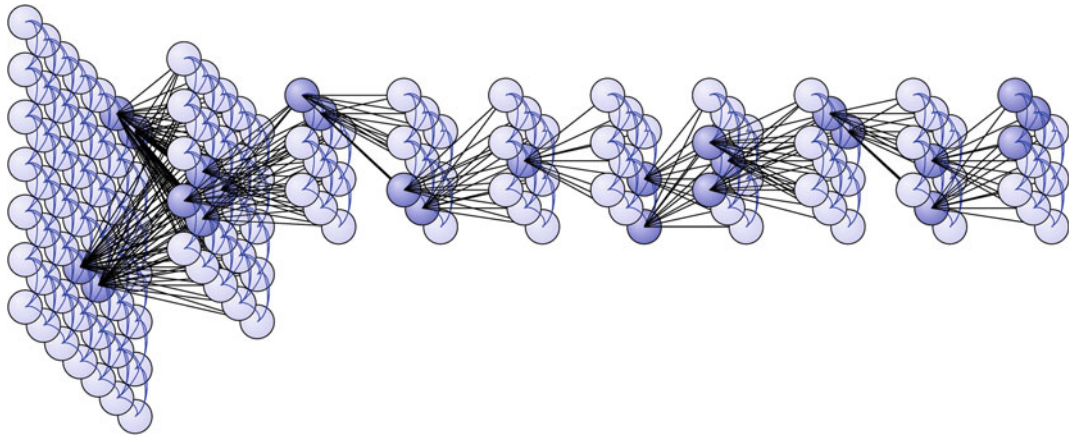


Fig. 2 miRNA-SOM classifier. *Dark blue* neurons have highly likely pre-miRNA candidates, which are input to the next level SOM (*black lines*)

Inputs :
G_ℓ: labeled input sequences (well-known pre-miRNAs)
G_u: unlabeled input sequences
n: initial map size ($n \times n$)
h_{max}: maximum deep level

Outputs:
C: pre-miRNA candidates at the last level

1 **begin**
2 $h \leftarrow 1$
3 $D_h \leftarrow G_u \cup G_\ell$
4 **while** $h < h_{max}$ **do**
5 Train a SOM with D_h
6 Label as miRNA-neuron those neurons having at least one sequence in G_ℓ
7 $h \leftarrow h+1$
8 $D_h \leftarrow$ sequences in miRNA-neurons
9 $C \leftarrow D_{h_{max}}$

Fig. 3 Unsupervised training and labeling of SOMs for novel pre-miRNA prediction from genome-wide data

are predicted as pre-miRNA candidates with a high probability of being miRNA precursors (line 9). This final list of top candidates is saved in the results folder of miRNA-SOM software. For practical applications of this model and the protocol, *see* **Notes 4** and **5**.

The deep structure of this classifier is shown in Fig. 2. When the root SOM, on the first layer, is trained and becomes stable, only the data in the neurons having clustered together with at least one well-known pre-miRNA are chosen as input data for training the next map, in the second layer. These neurons are marked miRNA-neurons and, although they might contain much more unlabeled data than labeled one, due to the existing high class-imbalance, they are marked as positive class neurons. During model training, only sequences clustered in miRNA-neurons remain for further training

the next deep level of miRNA-SOM. After training several layers, the best pre-miRNAs candidates are those sequences that remained in the miRNA-neurons at the last deep level.

With this approach, each internal map is trained only with a portion of the whole input genome-wide data. This method reduces significantly the number of possible candidate to pre-miRNAs, level after level, retaining at the last level only the high confidence candidates. In this last level, each well-known pre-miRNA in the miRNA-neurons (in dark blue) is grouped together with unlabeled sequences. They are selected as the best bona-fide candidates to novel pre-miRNAs.

4 Notes

1. In the first step (3.1), the recommended parameters for einverted are: gap penalty `$GAP = 6`; minimum score threshold `$THRESH = 25`; match score `$M = 3`; mismatch score `$MM = 3`; and maximum separation between the start and end of the inverted repeat `$DIST = 95`.
2. Also in the first step the recommended parameters to cut sequences are: window size `$WIN = 500000`; and window step `$step = 400000`.
3. It is recommended to start with a large initial SOM map, such as $n = 100$. After the first level, a large number of sequences will not pass to the next SOM level and they will be naturally discarded. After that, the map size number can be reduced.
4. A practical example on the application of this protocol to genome-wide data from *Echinococcus multilocularis* can be found in [13] and online in: http://fich.unl.edu.ar/sinc/web-demo/mirna-som/. The source code is available for free academic use at: http://sourceforge.net/projects/sourcesinc/files/mirnasom/ (download version 23).
5. Another example on a model organism (*Caenorhabditis elegans*) is available at: http://fich.unl.edu.ar/sinc/blog/web-demo/mirna-som-ce/.

References

1. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281–297
2. Esquela-Kerscher A, Slack FJ (2006) Oncomirs - microRNAs with a role in cancer. Nat Rev Cancer 6(1):259–269
3. Lecellier CH, Dunoyer P, Arar K, Lehmann-Che J, Eyquem S, Himber C, Saib A, Voinnet O (2005) A cellular MicroRNA mediates antiviral defense in human cells. Science 308 (5721):557–560
4. Rosenzvit M, Cucher M, Kamenetzky L, Macchiaroli N, Prada L, Camicia F (2013) MicroRNAs in endoparasites. Nova Science Publishers, New York
5. Li L, Xu J, Yang D, Tan X, Wang H (2010) Computational approaches for microRNA studies: a review. Mamm Genome 21(1):1–12

6. Lopes I de ON, Schliep A, de Carvalho A (2014) The discriminant power of RNA features for pre-miRNA recognition. BMC Bioinformatics 15(1):124+
7. Liu B, Li J, Cairns M (2014) Identifying mirnas, targets and functions. Brief Bioinform 15 (1):1–19
8. Lorenz R, Bernhart S, zu Siederdissen CH, Tafer H, Flamm C, Stadler P, Hofacker I (2011) ViennaRNA Package 2.0. Algorithms Mol Biol 6(1):26–36
9. Xue C, Li F, He T, Liu GP, Li Y, Zhang X (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinform 6(1):310
10. Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9(1):133–148
11. Yones C, Stegmayer G, Kamenetzky L, Milone D (2015) miRNAfe: a comprehensive tool for feature extraction in microRNA prediction. BioSystems 238:1–5
12. Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. J Mol Biol 215(1):403–410
13. Kamenetzky L, Stegmayer G, Maldonado L, Macchiaroli N, Yones C, Milone D (2016) MicroRNA discovery in the human parasite echinococcus multilocularis from genome-wide data. Genomics 107(6):274–280

Chapter 4

Protein Structure Modeling with MODELLER

Benjamin Webb and Andrej Sali

Abstract

Genome sequencing projects have resulted in a rapid increase in the number of known protein sequences. In contrast, only about one-hundredth of these sequences have been characterized at atomic resolution using experimental structure determination methods. Computational protein structure modeling techniques have the potential to bridge this sequence-structure gap. In the following chapter, we present an example that illustrates the use of MODELLER to construct a comparative model for a protein with unknown structure. Automation of a similar protocol has resulted in models of useful accuracy for domains in more than half of all known protein sequences.

Key words Comparative modeling, Fold assignment, Sequence-structure alignment, Model assessment, Multiple templates

1 Introduction

The function of a protein is determined by its sequence and its three-dimensional (3D) structure. Large-scale genome sequencing projects are providing researchers with millions of protein sequences, from various organisms, at an unprecedented pace. However, the rate of experimental structural characterization of these sequences is limited by the cost, time, and experimental challenges inherent in the structural determination by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy.

In the absence of experimentally determined structures, computationally derived protein structure models are often valuable for generating testable hypotheses [1, 2]. Such models are generally produced using either comparative modeling methods or free modeling techniques (also referred to as *ab initio* or *de novo* modeling) [3]. Comparative modeling relies on structural information from related proteins to guide the modeling procedure [4–6]. Free modeling does not require a related protein, but instead uses a variety of methods to combine physics with the known behaviors of protein structures (for example, by combining multiple short

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_4,

structural fragments extracted from known proteins) [7–9]; it is, however, extremely computationally expensive [3]. Comparative protein structure modeling, which this text focuses on, has been used to produce reliable structure models for at least one domain in more than half of all known sequences [10]. Hence, computational approaches can provide structural information for two orders of magnitude more sequences than experimental methods, and are expected to be increasingly relied upon as the gap between the number of known sequences and the number of experimentally determined structures continues to widen.

Comparative modeling consists of four main steps [4] (Fig. 1): (a) fold assignment that identifies overall similarity between the target sequence and at least one known structure (template); (b) alignment of the target sequence and the template(s); (c) building a model based on the alignment with the chosen template(s); and (d) predicting the accuracy of the model.

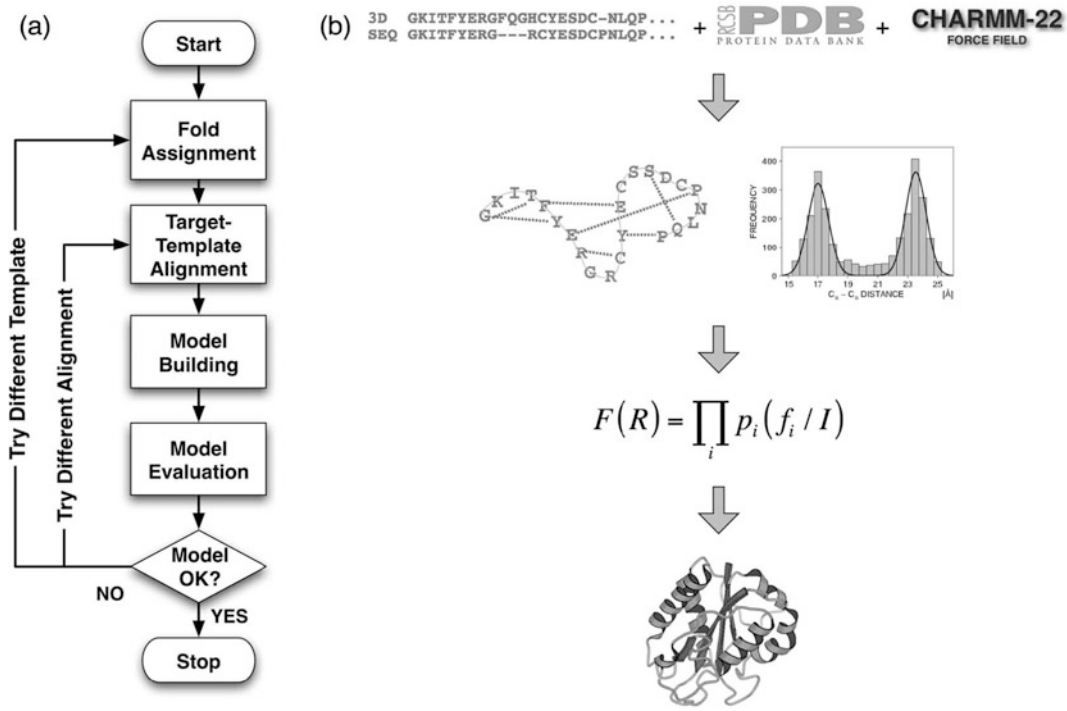


Fig. 1 Comparative protein structure modeling. (**a**) A flowchart illustrating the steps in the construction of a comparative model [4]. (**b**) Description of comparative modeling by extraction of spatial restraints as implemented in MODELLER [12]. By default, spatial restraints in MODELLER involve (1) homology-derived restraints from the aligned template structures, (2) statistical restraints derived from all known protein structures, and (3) stereochemical restraints from the CHARMM-22 molecular mechanics force-field. These restraints are combined into an objective function that is then optimized to calculate the final 3D model of the target sequence

MODELLER is a computer program for comparative protein structure modeling [11, 12]. In the simplest case, the input is an alignment of a sequence to be modeled with the template structure (s), the atomic coordinates of the template(s), and a simple script file. MODELLER then automatically calculates a model containing all non-hydrogen atoms, without any user intervention and within minutes on a desktop computer. Apart from model building, MODELLER can perform auxiliary tasks such as fold assignment, alignment of two protein sequences or their profiles [13], multiple alignment of protein sequences and/or structures [14, 15], clustering of sequences and/or structures, and *ab initio* modeling of loops in protein structures [11].

MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints that include (a) homology-derived restraints on the distances and dihedral angles in the target sequence, extracted from its alignment with the template structures [12], (b) stereochemical restraints such as bond length and bond angle preferences, obtained from the CHARMM-22 molecular mechanics force-field [16], (c) statistical preferences for dihedral angles and non-bonded inter-atomic distances, obtained from a representative set of known protein structures [17, 18], and (d) optional manually curated restraints, such as those from NMR spectroscopy, rules of secondary structure packing, cross-linking experiments, fluorescence spectroscopy, image reconstruction from electron microscopy, site-directed mutagenesis, and intuition (Fig. 1). The spatial restraints, expressed as probability density functions, are combined into an objective function that is optimized by a combination of conjugate gradients and molecular dynamics with simulated annealing. This model building procedure is similar to structure determination by NMR spectroscopy.

In this chapter, we use a sequence with unknown structure to illustrate the use of various modules in MODELLER to perform the four steps of comparative modeling.

2 Materials

To follow the examples in this discussion, both the MODELLER software and a set of suitable input files are needed. The MODELLER software is free for academic use; it can be downloaded from https://salilab.org/modeller/ and is available in binary form for most common machine types and operating systems (*see* **Note 1**). This text uses MODELLER 9.17, the most recent version at the time of writing, but the examples should also work with any newer version. The example input files can be downloaded from https://salilab.org/modeller/tutorial/FG17.zip.

All MODELLER scripts are Python scripts. Python is pre-installed on most Linux and Mac machines; Windows users can

obtain it from https://www.python.org/. It is not necessary to install Python, or to have a detailed knowledge of its use, to use MODELLER, but it is helpful for creating and understanding the more advanced MODELLER scripts.

2.1 Typographical Conventions

`Monospaced text` is used below for computer file and folder/directory names, command lines, file contents, and variable and class names.

3 Methods

The procedure for calculating a 3D model for a sequence with unknown structure will be illustrated using the following example: a novel gene for lactate dehydrogenase (LDH) was identified from the genomic sequence of Trichomonas vaginalis (TvLDH). The corresponding protein had higher sequence similarity to the malate dehydrogenase of the same species (TvMDH) than to any other LDH [19]. Comparative models were constructed for TvLDH and TvMDH to study the sequences in a structural context and to suggest site-directed mutagenesis experiments to elucidate changes in enzymatic specificity in this apparent case of convergent evolution. The native and mutated enzymes were subsequently expressed and their activities compared [19].

3.1 Fold Assignment

The first step in comparative modeling is to identify one or more templates (sequences with known 3D structure) for the modeling procedure. One way to do this is to search a database of experimentally determined structures extracted from the Protein Data Bank (PDB) [20] to find sequences that have detectable similarity to the target (*see* **Note 2**). To prepare this database (*see* **Note 3**), run the following command from the command line (*see* **Note 4**):

```
python make_pdb_95.py > make_pdb_95.log
```

This generates a file called `pdb_95.bin`, which is a binary representation of the search database (*see* **Note 5**) and a log file, `make_pdb_95.log`. Next, MODELLER's `profile.build()` command is used; this uses the local dynamic programming algorithm to identify sequences related to TvLDH [21]. In the simplest case, `profile.build()` takes as input the target sequence, in file `TvLDH.ali` (*see* **Note 6**), and the binary database and returns a set of statistically significant alignments (file `build_profile.prf`) and a MODELLER log file (`build_profile.log`). Run this step by typing

```
python build_profile.py > build_profile.log
```

The first few lines of the resulting `build_profile.prf` will look similar to (*see* **Note 7**) the following (note that the rightmost column, containing the primary sequence, has been omitted here for clarity):

```
# Number of sequences: 69
# Length of profile : 335
# N_PROF_ITERATIONS : 1
# GAP_PENALTIES_1D : -500.0 -50.0
# MATRIX_OFFSET : -450.0
# RR_FILE : ${LIB}/blosum62.sim.mat
 1 TvLDH S 0 335 1 335 0 0 0 0. 0.0
 2 1a5zA X 1 312 75 242 63 229 164 28. 0.58E-07
 3 2a92A X 1 316 8 191 6 186 174 26. 0.11E-03
 4 4aj2A X 1 327 85 301 89 300 207 25. 0.24E-04
 5 1b8pA X 1 327 7 331 6 325 316 42. 0.0
```

The first six lines of this file contain the input parameters used to create the alignments. Subsequent lines contain several columns of data; for the purposes of this example, the most important columns are (a) the second column, containing the PDB code of the related template sequences; (b) the eleventh column, containing the percentage sequence identity between the TvLDH and template sequences; and (c) the twelfth column, containing the *E*-values for the statistical significance of the alignments. These columns are shown in bold above.

The extent of similarity between the target-template pairs is usually quantified using sequence identity or a statistical measure such as *E*-value (*see* **Note 8**). Inspection of column 11 shows that a template with a high sequence identity with the target is the 1y7tA structure (45% sequence identity). Further inspection of column 12 shows that there are 14 PDB sequences, all but one corresponding to malate dehydrogenases (1b8pA, 1bdmA, 1civA, 3d5tA, 4h7pA, 4h7pB, 5mdhA, 7mdhA, 4tvoA, 4tvoB, 4uulA, 4uuoA, 4uupA, 1y7tA) that show significant similarities to TvLDH with *E*-values of zero.

3.2 Sequence-Structure Alignment

The next step is to align the target TvLDH sequence with the chosen template (*see* **Note 9**). Here, the 1y7tA template is used. This alignment is created using MODELLER's `align2d()` function (*see* **Note 10**). Although `align2d()` is based on a global dynamic programming algorithm [22], it is different from standard sequence-sequence alignment methods because it takes into account structural information from the template when constructing an alignment. This task is achieved through a variable gap penalty function that tends to place gaps in solvent exposed and curved regions, outside secondary structure segments, and not

between two positions that are close in space [14]. In the current example, the target-template similarity is so high that almost any method with reasonable parameters will result in the correct alignment (*see* **Note 11**).

This step is carried out by running:

```
python align2d.py > align2d.log
```

This script reads in the PDB structure of the template, and the sequence of the target (TvLDH) and calls the `align2d()` function to perform the alignment. The resulting alignment is written out in two formats. `TvLDH-1y7tA.ali` in the PIR format is subsequently used by MODELLER for modeling; `TvLDH-1y7tA.pap` in the PAP format is easier to read, for example, to see which residues are aligned with each other.

3.3 Model Building

Models of TvLDH can now be built by running:

```
python model.py > model.log
```

The script uses MODELLER's `automodel` class, specifying the name of the alignment file to use and the identifiers of the target (TvLDH) and template (1y7tA) sequences. It then asks `automodel` to generate five models (*see* **Note 12**). Each is assessed with the normalized DOPE assessment method [18]. The five models are written out as PDB files with names `TvLDH.B9999[0001-0005].pdb`.

3.4 Model Evaluation

The log file produced by the model building procedure (`model.log`) contains a summary of each calculation at the bottom of the file. This summary includes, for each of the five models, the MODELLER objective function (*see* **Note 13**) [12] and the normalized DOPE score (*see* **Note 14**). These scores can be used to identify which of the 5 models produced is likely to be the most accurate model (*see* **Note 15**).

Since the DOPE potential is simply a sum of interactions between pairs of atoms, it can be decomposed into a score per residue, which is termed in MODELLER an "energy profile." This energy profile can be generated for the model with the best DOPE score by running the `make_energy_profile.py` script. The script outputs the profile, `TvLDH.profile`, in a simple format that is easily displayed in any graphing package. Such a profile is useful to detect local regions of high pseudo-energy that usually correspond to errors in the model (*see* **Notes 16** and **17**).

3.5 Use of Multiple Templates

One way to potentially improve the accuracy of generated models is to use multiple template structures. When there are multiple templates, different template structures may be of higher local sequence

identity to the target (or higher quality) than others in different regions, allowing MODELLER to build a model based on the most useful structural information for each region in the protein. The procedure is demonstrated here using five templates that have high sequence identity to the target (1b8pA, 4h7pA, 4h7pB, 5mdhA, 1y7tA). Input files can be found in the "`multiple`" subdirectory of the zip-file. The first step is to align all of the templates with each other, which can be done by running:

```
python salign.py > salign.log
```

This script uses MODELLER's `salign()` function [15] to read in all of the template structures and then generate their best structural alignment (*see* **Note 18**), written out as `templates.ali`.

Next, just as for single template modeling, the target is aligned with the templates using the `align2d()` function. The function's `align_block` parameter is set to 5 to align the target sequence with the pre-aligned block of templates, and not to change the existing alignment between individual templates:

```
python align2d.py > align2d.log
```

Finally, model generation proceeds just as for the single template case (the only difference is that `automodel` is now given a list of all five templates):

```
python model.py > model.log
```

Comparison of the normalized DOPE scores from the end of this logfile with those from the single template case shows an improvement in the DOPE score of the best model from −0.92 to −1.19. Figure 2 shows the energy profiles of the best scoring models from each procedure (generated using the `plot_profiles.py` script). It can be seen that some of the predicted errors in the single template model (peaks in the graph) have been resolved in the model calculated using multiple templates.

3.6 External Assessment

Models generated by MODELLER are stored in PDB files, and so can be evaluated for accuracy with other methods if desired. One such method is the ModEval web server at https://salilab.org/evaluation/. This server takes as input the PDB file and the MODELLER PIR alignment used to generate it. It returns not only the normalized DOPE score and the energy profile, but also the GA341 assessment score [23, 24] and an estimate of the Cα RMSD and native overlap between the model and its hypothetical native structure, using the TSVMod method [25]; native overlap is

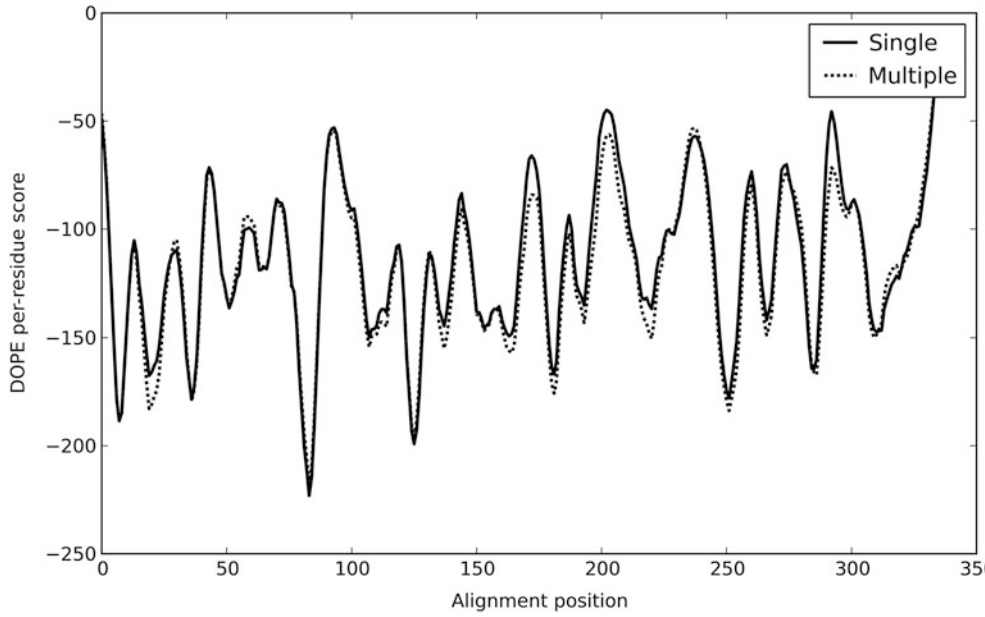


Fig. 2 The DOPE [18] energy profiles for the best-assessed model generated by modeling with a single template (*solid line*) and multiple templates (*dotted line*). Peaks (local regions of high, unfavorable score) tend to correspond to errors in the models

defined as the fraction of Cα atoms in the model that are within 3.5 Å of the same Cα atom in the native structure after least squares superposition.

3.7 Structures of Complexes

The example shown here generates a model of a single protein. However, MODELLER can also generate models of complexes of multiple proteins if templates for the entire complex are available; examples can be found in the MODELLER manual. In the case where only templates for the individual subunits in the complex can be found, comparative models can be docked in a pairwise fashion by molecular docking [26, 27] or assembled based on various experimental data to generate approximate models of the complex using a wide variety of integrative modeling methods [28–31]. For example, if a cryo-electron microscopy density map of the complex is available, a model of the whole complex can be constructed by simultaneously fitting comparative models of the subunits into the density map using the MultiFit method [32] or its associated web server at https://salilab.org/multifit/ [33]. Alternatively, if a small angle X-ray (SAXS) profile of a dimer is available, models of the dimer can be generated by docking the two subunits, constrained by the SAXS data, using the FoXSDock web server at https://salilab.org/foxsdock/ [34]. Both of these methods are part of the open source *Integrative Modeling Platform* (IMP) package [29].

4 Notes

1. The MODELLER website also contains a full manual, a mailing list, and more example MODELLER scripts. A license key is required to use MODELLER, but this can also be obtained from the website.
2. The sequence identity is a useful predictor of the accuracy of the final model when its value is >30%. It has been shown that models based on such alignments usually have, on average, more than ~60% of the backbone atoms correctly modeled with a root-mean-squared-deviation (RMSD) for Cα atoms of less than 3.5 Å (Fig. 3). Sequence-structure relationships in the "twilight zone" [35] (corresponding to relationships with

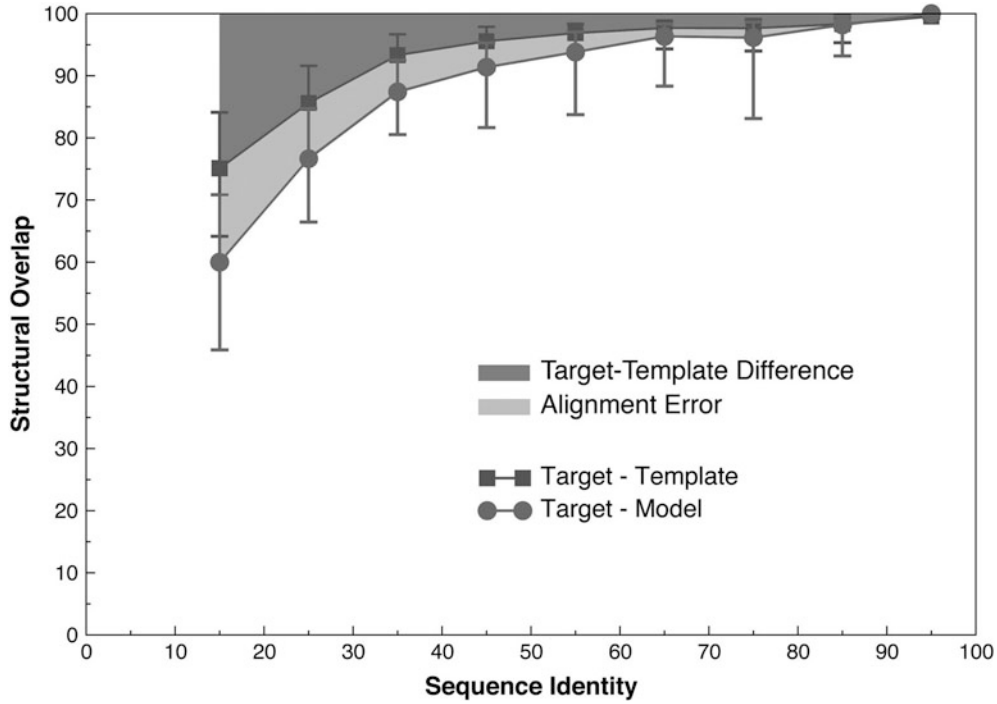


Fig. 3 Average model accuracy as a function of sequence identity [55]. As the sequence identity between the target sequence and the template structure decreases, the average structural similarity between the template and the target also decreases (*dark grey area*, *squares*) [56]. Structural overlap is defined as the fraction of equivalent Cα atoms. For the comparison of the model with the actual structure (*circles*), two Cα atoms were considered equivalent if they belonged to the same residue and were within 3.5 Å of each other after least squares superposition. For comparisons between the template structure and the actual target structure (*squares*), two Cα atoms were considered equivalent if they were within 3.5 Å of each other after alignment and rigid-body superposition. The difference between the model and the actual target structure is a combination of the target-template differences (*dark grey area*) and the alignment errors (*light grey area*). The figure was constructed by calculating ~1 million comparative models based on single template of varying similarity to the targets. All targets had known (experimentally determined) structures

statistically significant sequence similarity with identities generally in the 10–30% range), or the "midnight zone" [35] (corresponding to statistically insignificant sequence similarity), typically result in less accurate models.

3. The database contains sequences of the structures from PDB. To increase the search speed, redundancy is removed from the database; the PDB sequences are clustered with other sequences that are at least 95% identical, and only the representative of each cluster is stored in the database. This database is termed "pdb_95." A copy of this database is included in the downloaded zip-file as `pdb_95.pir`. Newer versions of this database, updated as new structures are deposited in PDB, can be downloaded from the MODELLER website at https://salilab.org/modeller/supplemental.html.
4. MODELLER is a command line tool, so all commands must be run by typing at the command line. All of the necessary input files for this demonstration are in the downloaded zip-file; simply download and extract the zip-file and change into the newly created directory (using the "`cd`" command at the command line). After this, MODELLER scripts can be run as shown in the text. All MODELLER scripts are Python scripts and so should be run with the "`python`" command. (On some systems the full path to the Python interpreter may be necessary, such as `/usr/bin/python` on a Linux or Mac machine or `C:\python27\python.exe` on a Windows system.) MODELLER scripts can also be run from other Python frontends, such as IDLE, if desired. On a Windows system, it is generally **not** a good idea to simply "double click" on a MODELLER Python script, since any output from the script will disappear as soon as it finishes. Finally, if Python is not installed, MODELLER includes a basic Python 2.3 interpreter as "`mod<version>`." For example, to run the first script using MODELLER version 9.17's own interpreter, run "`mod9.17 make_pdb_95.py`." Note that `mod9.17` automatically creates a "`make_pdb_95.log`" logfile.
5. The binary database is much faster to use than the original text format database, `pdb_95.pir`. Note, however, that it is not necessarily smaller. This script does not need to be run again unless `pdb_95.pir` is updated.
6. `TvLDH.ali` simply contains the primary sequence of the target, in MODELLER's variant of the PIR format (which is documented in more detail in the MODELLER manual). This file is included in the zip-file.
7. Although MODELLER's algorithms are deterministic, exactly the same job run on different machines (e.g., a Linux box *versus* a Windows or Mac machine) may give different results.

This difference may arise because different machines handle rounding of floating point numbers and ordering of floating point operations differently, and the minor differences introduced can be compounded and end up giving very different outputs. This variation is normal and to be expected, and so the results shown in this text may differ from those obtained by running MODELLER elsewhere.

8. The sequence identity is not a statistically reliable measure of alignment significance and corresponding model accuracy for values lower than 30% [35, 36]. During a scan of a large database, for instance, it is possible that low values occur purely by chance. In such cases, it is useful to quantify the sequence-structure relationship using more robust measures of statistical significance, such as *E*-values [37], that compare the score obtained for an alignment with an established background distribution of such scores.

   One other problem of using sequence identity as a measure to select templates is that, in practice, there is no single generally used way to normalize it [36]. For instance, local alignment methods usually normalize the number of identically aligned residues by the length of the alignment, while global alignment methods normalize it by either the length of the target sequence or the length of the shorter of the two sequences. Therefore, it is possible that alignments of short fragments produce a high sequence identity but do not result in an accurate model. Measures of statistical significance do not suffer from this normalization problem because the alignment scores are corrected for the length of the aligned segment before the significance is computed [37, 38].
9. After a list of all related protein structures and their alignments with the target sequence has been obtained, template structures are usually prioritized depending on the purpose of the comparative model. Template structures may be chosen based purely on the target-template sequence identity or a combination of several other criteria, such as the experimental accuracy of the structures (resolution of X-ray structures, number of restraints per residue for NMR structures), conservation of active-site residues, holo-structures that have bound ligands of interest, and prior biological information that pertains to the solvent, *p*H, and quaternary contacts. In this case an MDH template with a moderately high sequence identity was chosen. (In practice, the modeling can be simply repeated with a different template or set of templates and the resulting models compared for utility.) One of the detected templates, 4uulA, is TvLDH itself, the structure of which was recently determined in a study of convergent evolution of LDH and MDH [39]; this template was excluded from selection in order to demonstrate the comparative modeling method.

10. Although fold assignment and sequence-structure alignment are logically two distinct steps in the process of comparative modeling, in practice almost all fold assignment methods also provide sequence-structure alignments. In the past, fold assignment methods were optimized for better sensitivity in detecting remotely related homologs, often at the cost of alignment accuracy. However, recent methods simultaneously optimize both the sensitivity and alignment accuracy. For the sake of clarity, however, they are still considered as separate steps in the current chapter.
11. Most alignment methods use either the local or global dynamic programming algorithms to derive the optimal alignment between two or more sequences and/or structures. The methods, however, vary in terms of the scoring function that is being optimized. The differences are usually in the form of the gap penalty function (linear, affine, or variable) [14], the substitution matrix used to score the aligned residues (20×20 matrices derived from alignments with a given sequence identity, those derived from structural alignments, and those incorporating the structural environment of the residues) [40], or combinations of both [41–44]. There doesn't yet exist a single universal scoring function that guarantees the most accurate alignment for all situations. Above 30–40% sequence identity, alignments produced by almost all methods are similar. However, in the twilight and midnight zones of sequence identity, models based on the alignments of different methods tend to have significant variations in accuracy. Improving the performance and accuracy of methods in this regime remains one of the main tasks of comparative modeling [45, 46].
12. To generate each model, MODELLER takes a starting structure, which is simply the target sequence threaded onto the template backbone, adds some randomization to the coordinates, and then optimizes it by searching for the minimum of its scoring function. Since finding the global minimum of the scoring function is not guaranteed, it is usually recommended to repeat the procedure multiple times to generate an ensemble of models; the randomization is necessary, otherwise the same model would be generated each time. Computing multiple models is particularly important when the sequence-structure alignment contains different templates with many insertions and/or deletions. Calculating multiple models allows for better sampling of the different template segments and the conformations of the unaligned regions. The best scoring model among these multiple models is generally more accurate than the first model produced.
13. The MODELLER objective function is a measure of how well the model satisfies the input spatial restraints. Lower values of

the objective function indicate a better fit with the input data and, thus, models that are likely to be more accurate [12].

14. The Discrete Optimized Protein Energy (DOPE) [18] is an atomic distance-dependent statistical potential based on a physical reference state that accounts for the finite size and spherical shape of proteins. The reference state assumes that a protein chain consists of non-interacting atoms in a homogeneous sphere of equivalent radius to that of the corresponding protein. The DOPE potential was derived by comparing the distance statistics from a non-redundant PDB subset of 1472 high-resolution protein structures with the distance distribution function of the reference state. By default, the DOPE score is not included in the model building routine, and thus can be used as an independent assessment of the accuracy of the output models. The DOPE score assigns a score for a model by considering the positions of all non-hydrogen atoms, with lower scores predicting more accurate models. Since DOPE is a pseudo-energy dependent on the composition and size of the system, DOPE scores are only directly comparable for models with the same set of atoms (so can, for example, be used to rank multiple models of the same protein, but cannot be used without additional approximations to compare models of a protein and its mutant). The normalized DOPE (or z-DOPE) score, however, is a z score that relates the DOPE score of the model to the average observed DOPE score for "reference" protein structures of similar size [25]. Negative normalized DOPE scores of -1 or below are likely to correspond to models with the correct fold.
15. Different measures to predict errors in a protein structure perform best at different levels of resolution. For instance, physics-based force-fields may be helpful at identifying the best model when all models are very close to the native state (<1.5 Å RMSD, corresponding to ~85% target-template sequence identity). In contrast, coarse-grained scores such as atomic distance statistical potentials have been shown to have the greatest ability to differentiate models in the ~3 Å Cα RMSD range. Tests show that such scores are often able to identify a model within 0.5 Å Cα RMSD of the most accurate model produced [47]. When multiple models are built, the DOPE score generally selects a more accurate model than the MODELLER objective function.
16. Segments of the target sequence that have no equivalent region in the template structure (*i.e.*, insertions or loops) are among the most difficult regions to model [11, 48–50]. This difficulty is compounded when the target and template are distantly related, with errors in the alignment leading to incorrect positions of the insertions and distortions in the loop environment.

Using alignment methods that incorporate structural information can often correct such errors [14]. Once a reliable alignment is obtained, various modeling protocols can predict the loop conformation, for insertions of less than approximately 15 residues long [11, 48, 51, 52].

17. As a consequence of sequence divergence, the mainchain conformation of a protein can change, even if the overall fold remains the same. Therefore, it is possible that in some correctly aligned segments of a model, the template is locally different (<3 Å) from the target, resulting in errors in that region. The structural differences are sometimes not due to differences in sequence, but are a consequence of artifacts in structure determination or structure determination in different environments (e.g., packing of subunits in a crystal and ligands). The simultaneous use of several templates can minimize this kind of error [53, 54].
18. It is particularly important to generate the best alignment of the structures to minimize conflicting information (e.g., one template suggesting that two Cα atoms in the target are close, and another suggesting they are widely separated). SALIGN [15] uses both sequence- and structure-dependent features to align multiple structures. It employs an iterative procedure to determine the input parameters that maximize the structural overlap of the generated alignment.

Acknowledgements

We are grateful to all members of our research group. We also acknowledge support from National Institutes of Health (U54 GM094625) as well as computing hardware support from Ron Conway, Mike Homer, Hewlett-Packard, NetApp, IBM, and Intel.

References

1. Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294 (5540):93–96
2. Schwede T, Sali A, Honig B, Levitt M, Berman H, Jones D, Brenner S, Burley S, Das R, Dokholyan N, Dunbrack RJ, Fidelis K, Fiser A, Godzik A, Huang Y, Humblet C, Jacobson M, Joachimiak A, Krystek SJ, Kortemme T, Kryshtafovych A, Montelione G, Moult J, Murray D, Sanchez R, Sosnick T, Standley D, Stouch T, Vajda S, Vasquez M, Westbrook J, Wilson I (2009) Outcome of a workshop on applications of protein models in biomedical research. Structure 17(2):151–159
3. Zhang Y (2008) Progress and challenges in protein structure prediction. Curr Opin Struct Biol 18(3):342–348. doi:10.1016/j.sbi.2008.02.004
4. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325
5. Eswar N, Sali A (2009) Protein structure modeling. In: Sussman JL, Spadon P (eds) From molecules to medicine, structure of biological macromolecules and its relevance in combating new diseases and bioterrorism. NATO Science for peace and security series –

A: chemistry and biology. Springer-Verlag, Dordrecht, pp 139–151
6. Ginalski K (2006) Comparative modeling for protein structure prediction. Curr Opin Struct Biol 16(2):172–177. doi:10.1016/j.sbi.2006.02.003
7. Das R, Baker D (2008) Macromolecular modeling with rosetta. Annu Rev Biochem 77:363–382. doi:10.1146/annurev.biochem.77.062906.171838
8. Zhang Y, Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 101(20):7594–7599. doi:10.1073/pnas.0305695101
9. Simons KT, Bonneau R, Ruczinski I, Baker D (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins Suppl 3:171–176
10. Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, Braberg H, Yang Z, Meng EC, Pettersen EF, Huang CC, Datta RS, Sampathkumar P, Madhusudhan MS, Sjolander K, Ferrin TE, Burley SK, Sali A (2011) ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 39:465–474
11. Fiser A, Do RKG, Sali A (2000) Modeling of loops in protein structures. Protein Sci 9 (9):1753–1773
12. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815
13. Marti-Renom MA, Madhusudhan MS, Sali A (2004) Alignment of protein sequences by their profiles. Protein Sci 13(4):1071–1087
14. Madhusudhan MS, Marti-Renom MA, Sanchez R, Sali A (2006) Variable gap penalty for protein sequence-structure alignment. Protein Eng Des Sel 19(3):129–133
15. Madhusudhan MS, Webb BM, Marti-Renom MA, Eswar N, Sali A (2009) Alignment of multiple protein structures based on sequence and structure features. Protein Eng Des Sel 22:569–574
16. Brooks BR, Brooks CL 3rd, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30 (10):1545–1614. doi:10.1002/jcc.21287
17. Sali A, Overington JP (1994) Derivation of rules for comparative protein modeling from a database of protein structure alignments. Protein Sci 3(9):1582–1596
18. Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15(11):2507–2524
19. Wu G, Fiser A, ter Kuile B, Sali A, Muller M (1999) Convergent evolution of Trichomonas vaginalis lactate dehydrogenase from malate dehydrogenase. Proc Natl Acad Sci USA 96 (11):6285–6290
20. 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
21. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147(1):195–197
22. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453
23. John B, Sali A (2003) Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 31(14):3982–3992. doi:10.1093/nar/gkg460
24. Melo F, Sanchez R, Sali A (2002) Statistical potentials for fold assessment. Protein Sci 11 (2):430–448. doi:10.1110/ps.22802
25. Eramian D, Eswar N, Shen M, Sali A (2008) How well can the accuracy of comparative protein structure models be predicted? Protein Sci 17(11):1881–1893
26. Vajda S, Kozakov D (2009) Convergence and combination of methods in protein-protein docking. Curr Opin Struct Biol 19 (2):164–170. doi:10.1016/j.sbi.2009.02.008
27. Lensink MF, Wodak SJ (2010) Docking and scoring protein interactions: CAPRI 2009. Proteins 78(15):3073–3084. doi:10.1002/prot.22818
28. Alber F, Forster F, Korkin D, Topf M, Sali A (2008) Integrating diverse data for structure determination of macromolecular assemblies. Annu Rev Biochem 77:443–477
29. Russel D, Lasker K, Webb B, Velazquez-Muriel J, Tjioe E, Schneidman-Duhovny D, Peterson B, Sali A (2012) Putting the pieces together: integrative structure determination of macromolecular assemblies. PLoS Biol 10(1): e1001244
30. Robinson C, Sali A, Baumeister W (2007) The molecular sociology of the cell. Nature 450 (7172):973–982

31. Ward A, Sali A, Wilson I (2013) Integrative structural biology. Science 339 (6122):913–915
32. Lasker K, Sali A, Wolfson HJ (2010) Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. Proteins:Struct Funct Bioinform 78:3205–3211
33. Tjioe E, Lasker K, Webb B, Wolfson H, Sali A (2011) MultiFit: a web server for fitting multiple protein structures into their electron microscopy density map. Nucleic Acids Res 39:167–170
34. Schneidman-Duhovny D, Hammel M, Sali A (2011) Macromolecular docking restrained by a small angle X-ray scattering profile. J Struct Biol 3:461–471
35. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12(2):85–94
36. May AC (2004) Percent sequence identity; the need to be explicit. Structure 12(5):737–738. doi:10.1016/j.str.2004.04.001
37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402
38. Pearson WR (1998) Empirical statistical estimates for sequence similarity searches. J Mol Biol 276(1):71–84. doi:10.1006/jmbi.1997.1525
39. Steindel PA, Chen EH, Wirth JD, Theobald DL (2016) Gradual neofunctionalization in the convergent evolution of trichomonad lactate and malate dehydrogenases. Protein Sci 25 (7):1319–1331. doi:10.1002/pro.2904
40. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89(22):10915–10919
41. Zhou H, Zhou Y (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 58 (2):321–328. doi:10.1002/prot.20308
42. McGuffin LJ, Jones DT (2003) Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics 19 (7):874–881
43. Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins 51 (4):504–514. doi:10.1002/prot.10369
44. Shi J, Blundell TL, Mizuguchi K (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310(1):243–257. doi:10.1006/jmbi.2001.4762
45. Dunbrack RL Jr (2006) Sequence comparison and protein structure prediction. Curr Opin Struct Biol 16(3):374–384. doi:10.1016/j.sbi.2006.05.006
46. Xiang Z (2006) Advances in homology protein structure modeling. Curr Protein Pept Sci 7 (3):217–227
47. Eramian D, Shen M, Devos D, Melo F, Sali A, Marti-Renom M (2006) A composite score for predicting errors in protein structure models. Protein Sci 15(7):1653–1666
48. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA (2004) A hierarchical approach to all-atom protein loop prediction. Proteins 55(2):351–367. doi:10.1002/prot.10613
49. Zhao S, Zhu K, Li J, Friesner RA (2011) Progress in super long loop prediction. Proteins 79 (10):2920–2935. doi:10.1002/prot.23129
50. Fernandez-Fuentes N, Oliva B, Fiser A (2006) A supersecondary structure library and search algorithm for modeling loops in protein structures. Nucleic Acids Res 34(7):2085–2097. doi:10.1093/nar/gkl156
51. van Vlijmen HW, Karplus M (1997) PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 267(4):975–1001. doi:10.1006/jmbi.1996.0857
52. Coutsias EA, Seok C, Jacobson MP, Dill KA (2004) A kinematic view of loop closure. J Comput Chem 25(4):510–528. doi:10.1002/jcc.10416
53. Sanchez R, Sali A (1997) Evaluation of comparative protein structure modeling by MODELLER-3. Proteins Suppl 1:50–58
54. Srinivasan N, Blundell TL (1993) An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. Protein Eng 6(5):501–512
55. Sanchez R, Sali A (1998) Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proc Natl Acad Sci USA 95 (23):13597–13602
56. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. Embo J 5(4):823–826

Chapter 5

Protein Function Prediction

Leonardo Magalhães Cruz, Sheyla Trefflich, Vinícius Almir Weiss, and Mauro Antônio Alves Castro

Abstract

Protein function is a concept that can have different interpretations in different biological contexts, and the number and diversity of novel proteins identified by large-scale "omics" technologies poses increasingly new challenges. In this review we explore current strategies used to predict protein function focused on high-throughput sequence analysis, as for example, inference based on sequence similarity, sequence composition, structure, and protein–protein interaction. Various prediction strategies are discussed together with illustrative workflows highlighting the use of some benchmark tools and knowledge bases in the field.

Key words Protein function, Homology, Ontology, Biological databases, Database sequence similarity search, Protein families, Protein domains, Phylogeny, Bioinformatics

1 Introduction

With the advent of structural, functional, and comparative genomics, numerous sequences of predicted proteins have been produced in a velocity that cannot be followed by its experimental studies, and the only feasible way to annotate tentative functions to these proteins is by means of automatic sequence analysis [1]. Beyond sequence, structural genomic projects have also allowed the determination of protein structure in a high-throughput fashion [2]. On the other hand, although these methodologies contribute to our knowledge, over one-third of structures are of proteins of unknown function and their worth can only be significantly enhanced by knowing the biological roles that they play [2], but experimental characterization of function cannot scale up to accommodate the vast amount of sequence [3] and structural data already available and the growing gap between sequences and experimentally annotated proteins can only be accomplished by combining experimental and computational methods for functional annotation [4]. Further, experimental efforts have been

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_5,

done to determine protein function and provide a more detailed understanding mainly of model organisms, expecting that accurate annotation may be transferred to other species by computational methods [4]. Numerous approaches have been used to automatically predict protein function so far, from different data types, such as sequence information, protein structure, phylogenetics and evolutionary relationships, interaction and association data, and a combination of these [5].

The accurate annotation of protein function is a key to understanding life at the molecular level and has great biomedical and pharmaceutical implications [3, 4]. In the absence of experimental data, the function of a protein can be inferred on the basis of its sequence similarity, sequence composition, structure [3], gene expression, protein–protein interaction, phylogeny, genomic context, or other structural or functional information based on our knowledge about proteins with already known functions. Even in the presence of some experimental evidences, automatic analysis is important to integrate data and evidences for function, because experimental characterization of a protein such as structural data, analysis of gene expression, and delineation of a protein interaction network rarely gives direct clues to gene function [6]. The computational annotation of protein function has therefore emerged as a problem at the forefront of computational and molecular biology [3]. However, prediction of function from sequence is a considerably more complex enterprise than a simple sequence database search [7].

1.1 Homology

Similar genes often have conserved functions in different organisms. This happens because organisms share a common evolutionary history, preserving functions from a common ancestor and changing it along time of evolution. These shared functions or characteristics linked by a common ancestor is called homology and cannot be quantified. Functions or characteristics "are" or "are not" homologous. However, ancestral organisms or states are not present today and the way to infer homology is by means of quantification of similarity. For nucleotide and amino acid sequences, the way to measure similarity is by means of a sequence alignment. Different types of homologies may be distinguished and the main ones are (Fig. 1a): orthologs, arisen by an speciation evolutionary event, and paralogs, arisen by a gene duplication evolutionary event. Time of evolution may modify nucleotide or protein sequences and lengths, but it is also important to consider the evolution of proteins in another perspective. Many proteins are structured in a domain manner, meaning that these proteins are composed by a set of independent functional units and each of these domains may have a different evolutionary history.

The more the organisms evolve, the more the sequences diverge and the more difficult is it to establish similarity and infer homology from similarity and sequence alignments. This relationship can also

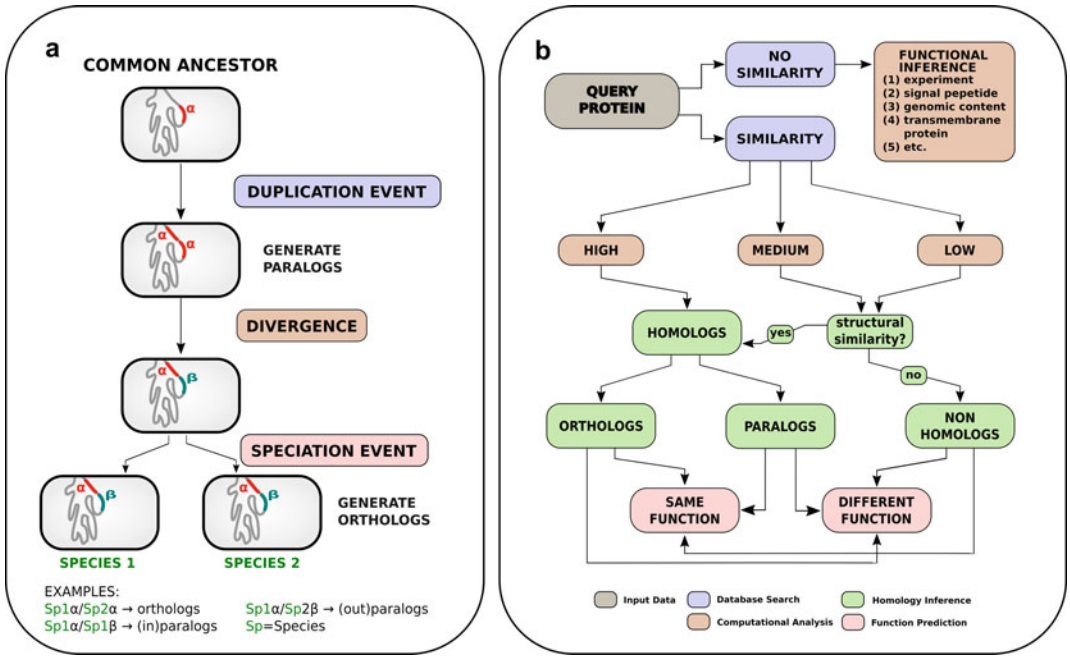


Fig. 1 Sequence similarity and homology in protein function prediction. Flowcharts summarizing (**a**) basic concepts on homology and sequence divergence and (**b**) possible strategies in protein annotation using sequence similarity

be used for protein function prediction, where the higher the sequence similarity, the better the chance that homologous proteins in fact share functional features [8]. For that purpose, the following rule may be useful [9]: (a) 90% of protein sequences sharing 30%, or more, identity are structurally similar, suggesting high probability of homology and also function; (b) only 10% of protein sequences sharing 25%, or less, sequence identity are structurally similar, suggesting a low probability to find homology and function.

Directly or indirectly, the prediction of a protein function in silico passes through the identification of homologous and the measurement of similarity that, at the end, will allow homology identification. On the contrary, displacement of non-homologous but functionally equivalent enzymes [7] is also observed.

1.2 Definition of Protein Function

Protein function is a concept that can have different interpretations in different biological contexts and/or level [8, 10–12], describing biochemical, cellular, and phenotypic aspects of the molecular events that involve the protein [3, 4]. The protein function can be divided into three major categories: (a) molecular function, e.g., the specific reaction catalyzed by an enzyme; (b) biological process, e.g., the metabolic pathway the enzyme is involved in; and (c) system or physiological level, e.g., if the enzyme is involved in respiration, photosynthesis, cell signaling, etc. One could also

consider a fourth level of cellular component, specifying the compartment of the cell the protein plays its role, e.g., cell membrane, any organelles [8, 11–13]. Protein function may also vary in space and time [11], as we will see, for example, in the case of moonlighting proteins. Computational methods exist to predict all of these aspects of function [13]. Furthermore, most biological processes are carried out by groups of interacting proteins and these interactions can be predicted in silico [13]. These many levels of protein function, from a very specific biochemical activity to a biological processes and pathways context, and from the cell to the organism level [2] generate practical consequences with protein annotation including vague terms to describe its function, such as "like protein," "containing domain protein," and "signaling protein" [2].

When attempting to identify the molecular function of a protein, it is important to bear in mind the simple rule: sequence → structure → function, that is, sequence determines the structure and structure determines the molecular function.

When describing function, attention must be paid to two kinds of proteins: those containing multiple domains and the so calling moonlighting proteins. The former are proteins composed of many domains, each domain contributing with a different specialized function to compose a unique biological function of the protein. Variation in the domain composition may occur, given different functions to similar proteins within the same family. The last are proteins that perform more than one function (multitask protein). For a moonlighting protein, usually independent unrelated functions are observed [14], not including function variation that results from gene fusions, homologous but nonidentical proteins, proteins resulting from alternative splicing, variation in posttranslational modifications and proteins operating in different locations or are able to utilize different substrates but have a single function [15].

It is now recognized that multifunctional proteins are common [4]. At least 34% of functionally characterized proteins (by experimental studies) are already assigned more than one distinct molecular function term and that at least 56% of proteins participate in more than one distinct biological process [4].

Different function of moonlighting proteins occur due to [15]: (a) cellular localization (within the cell or if inside/outside the cell); (b) the cell types expressing the protein; (c) the substrate, product, or a cofactor bound to the protein or different binding sites for different ligands; (d) the number of subunits joined and variation in the complexes to form the quaternary structure of a protein. These mechanisms that a protein can moonlight demonstrate the function may shift at different levels (i.e., molecular function, cellular process, or localization). The MoonProt database actually lists approximately 300 experimentally identified moonlighting proteins (www.moonlightingproteins.org).

If the moonlighting functions of a protein may also be assigned to an unknown protein by means of homology-based transfer is a matter of discussion. Identification of additional function of moonlight proteins is relatively recent and difficult by experimentation and its identification by in silico analysis is an even greater challenge [14]. Few methods are actually available to predict moonlighting proteins. Khan et al. [14] searched GO for known moonlighting proteins and observed that clusters of these proteins reflect their functions. Further analysis of protein–protein interaction, gene expression, phylogenetic profile, and genetic interaction network revealed that moonlighting proteins physically interact with a higher number of distinct functional classes of proteins than non-moonlighting proteins and that moonlighting proteins tend to interact with other moonlighting proteins. It has also been suggested that moonlighting proteins are under positive selection [14, 15]. These observations open the door for in silico prediction of moonlighting functions.

1.3 Proteins of Unknown Function

A large portion of known proteins are poorly characterized experimentally, with very little knowledge about their function [8]. The vast majority of proteins with function experimentally verified is observed in model organisms [4], but even for those organisms, a significant part of all proteins coded in their genomes are to be characterized. In *Escherichia coli* K-12, about one-third (1408) of the 4225 predicted proteins remain functionally unannotated (orphans) and only half of the predicted proteins have indicative of function based on experimental evidence and the same proportion seems to apply to *Saccharomyces cerevisiae* [6, 16]. Further, the remaining genes between experimentally annotated and unannotated in *E. coli* have either only generic functional attributes [16].

In Swiss-Prot v15.15, a curated database, approximately 90% of annotated proteins in Molecular Function and Biological Process ontologies belong to nine model organisms only (*H. sapiens*, *S. cerevisiae*, *M. musculus*, *R. norvegicus*, *A. thaliana*, *D. melanogaster*, *S. pombe*, *E. coli* K-12, and *C. elegans*) [4]. However, nearly 60% of the proteins from these model organisms still do not have any experimentally determined Molecular Function or Biological Process terms [4].

In CharProtDB (www.jcvi.org/charprotdb) [17], a database of experimentally characterized proteins, updated dataset till 2011 indicate that the main organisms with experimentally characterized proteins are as follow: *Escherichia coli* with 2631 proteins (~60% of all proteins), *Schizosaccharomyces pombe* with 1817 proteins (~35%), *Candida albicans* with 1308 proteins (~9%), and *Bacillus subtillis* with 1250 proteins (~30%). A total of 1252 species of all domain of life are included in the database and 96% of them have less than 100 experimentally characterized proteins.

Although these information about experimentally characterized proteins is difficult to obtain and is presented from different source and time, taken together, they give us an overview of our current knowledge about the function of proteins in different organisms and our need for tools that allow of automatic and reliable prediction of protein function.

In Pfam (pfam.xfam.org) [18] release 26.0, a database dedicated to protein families and its domains, more than 20% of all proteins are annotated as containing DUFs (Domains of Unknown Function) [19]. A total of 355 essential proteins in 16 model bacterial species contain 238 DUFs, most of which represent single-domain proteins, clearly establishing the biological essentiality of DUFs [19]. About 9% of DUFs spanned all domains of life, nearly half (43%) had been detected only in bacteria, 19% were only found in eukaryotes, and 3% are restricted to Archaea [20].

For the updated version of COG (Clusters of Orthologous Groups; www.ncbi.nlm.nih.gov/COG) [1], a database of putative orthologous proteins shared from completely sequenced genomes of bacteria and archaea, among a total of 4631 COGs distributed in 26 functional categories, R "General function prediction only" (507 COGs) and S "Function unknown" (959 COGs) are the most abundant categories, both counting for 31.6% of all COGs. Further, all COGs include about 60% and 86% of bacterial and archaeal proteomes, respectively [1], with remaining proteins not even being assigned to any existing COG. The fraction of the total proteome with specific functional annotation (excluding R and S categories) varies from a minimum of about 51–53% to a maximum of 72–76% at the phyla level [1].

The large number of functionally unannotated genes is observed because experimental characterization is time consuming, so these genes have never been studied experimentally or experimental studies brought contradictory results that could not be easily reconciled [6].

2 Strategies for Protein Function Prediction

Normally, the prediction of a protein function starts by trying to define its molecular function, using a homology-based transfer strategy, e.g., a similarity search against a database of known proteins or a search against a protein family and domain database. In a next step, one tries to extend the molecular function to a system function, that is, define the role played by a protein in a biological process.

Computational biology offers tools that can provide insight into the function of proteins based on their sequence, their structure, their evolutionary history, and their association with other proteins [8]. There are also methods that directly analyze the

sequence or structure in order to predict the function or methods that rely on sources of information that are beyond the protein itself, such as genomic context, protein–protein interaction networks, or membership in biochemical pathways [8].

Prediction of protein function, unlike establishing homology, is not a "yes" or "no" decision (i.e., an unknown protein will or will not have exactly the same function than a homologous counterpart). Function may be shared at different levels. The obvious example is two proteins that participate in the same cellular process but have different enzymatic activities (i.e., share the same cellular process function but have different molecular functions). Further, if two proteins are homologous, it means that they share a common evolutionary origin, but it does not guarantee that these two proteins will have the same function [8]. On the other hand, concerning about different kinds of homology, in general, functions from ancestral origin tend to be conserved more in orthologs than in paralogs [8, 21], but frequently distinguishing between them is not a straightforward task and even orthologs may diverge functionally [8, 21]. In the opposite way, proteins with same function may arise not by means of homology, but by convergent evolution, when by means of adaptive change, some molecular "functionality" arises independently in proteins not sharing an ancestral sequence [22, 23]. All these possibilities are presented in Fig. 1b, showing how homology, similarity, and function correlate.

Function predicted automatically and on a large scale includes additional problems concerning the need to standardize and quantitatively assess the similarity of functions between proteins [8]. A large number of methods have been proposed to predict protein function using information from amino acid sequence and predicted physicochemical properties, phylogenetic profiles and genomic context, protein–protein interaction networks, protein structure data, microarrays and clustering patterns of coregulated genes, predicted ligands, or a combination of data types [3, 4, 24].

The primary databases of biological sequences and structures are the main sources of information for any methods attempting to predict protein function. These databases can be directly searched to looking for similar sequences or structures and infer homology to transfer functional annotation or can be used to build secondary databases of clusters of protein sequences (e.g., COG, UniProtKB/UniRef, NCBI Protein Clusters, Panther), family and domains (e.g., Pfam, PROSITE, SMART, PRINTS, CDD), protein domain classification from structures and sequences (e.g., CATH, Gene3D), or retrieve well-known and annotated sequences/structures experimentally characterized to build probabilistic models or models based on machine learning that may be applied to scan unknown proteins to give insight in its function (e.g., TMHMM, LocTree3, BaCelLo, TargetP, PSORT, Protein prowler, LipoP, TatP). In this sense, all knowledge applied to automatically predict

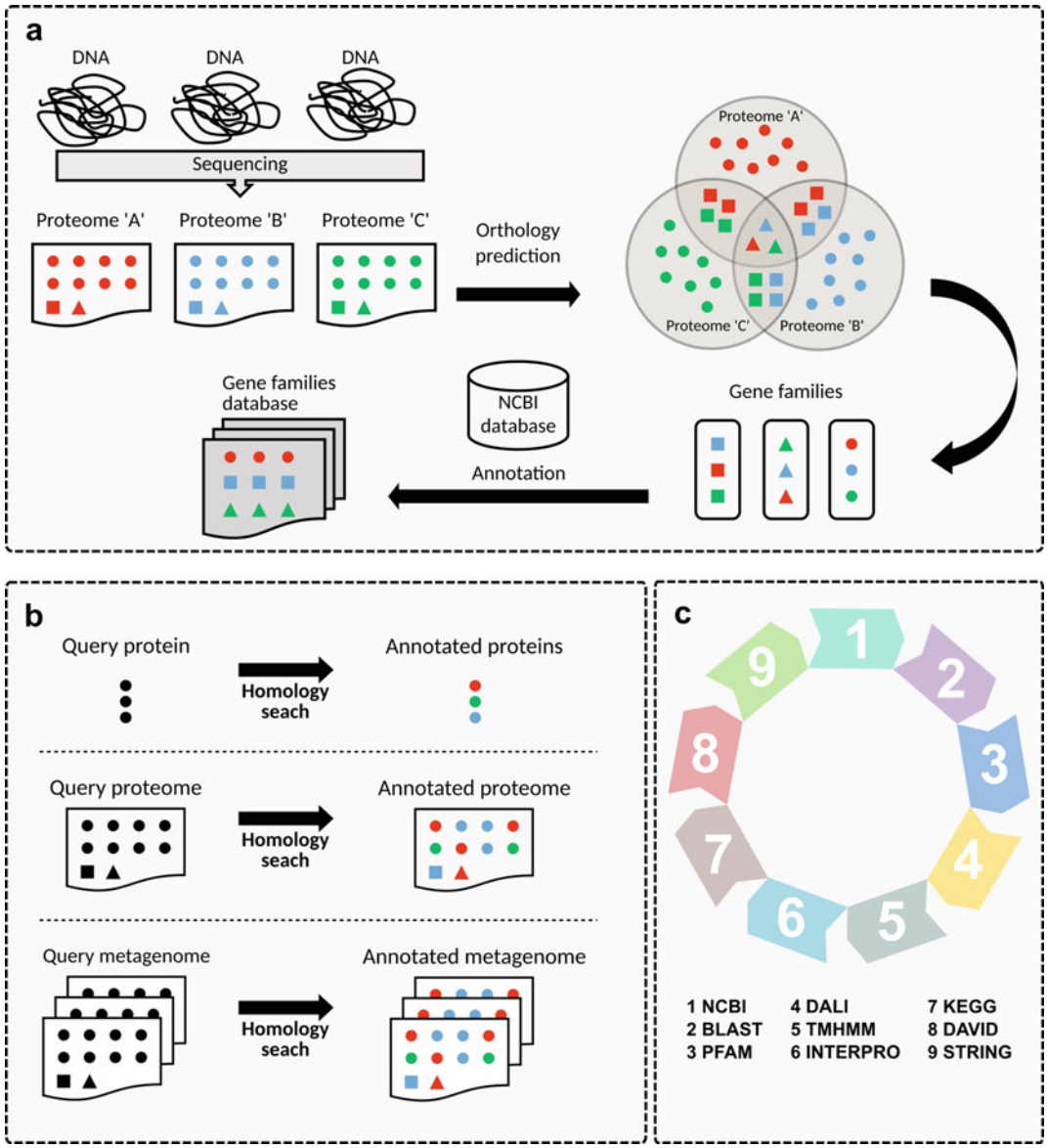


Fig. 2 Protein annotation strategies using knowledge bases. Flowcharts exemplifying (**a**) knowledge base construction and (**b**) the annotation process of a protein sequence, a proteome and a metagenome using homology-searching strategies. (**c**) The combination of different resources can be used for knowledge discovery in databases in order to help the annotation process (*see* Fig. 3 for additional details)

the function of a protein from its sequence and/or structure is founded on the concept of homology and in the known proteins and annotation deposited in the databases, that is, the automatic prediction will use these information directly or indirectly. An example showing the steps of some of these databases may be built is presented in Fig. 2a, starting from DNA sequencing, generally producing complete genome sequences, to the knowledge

database, passing through identification of orthologs, clustering sequences in gene families, and automatic and manual annotations. This knowledge is then used to predict function from single proteins, complete proteomes or even metaproteomes (Fig. 2b) using many available bioinformatic tools applying different methodologies (Fig. 2c) as outlined below and detailed in Fig. 3, including commonly used tools with a simplified workflow of analysis.

2.1 Sequence-Based Methods

2.1.1 Sequence Similarity/Homology-Based Transfer

Currently, the simplest and most used method to determine protein function is based on similarity search. This is accomplished by means of similarity search programs, with BLAST (blast.ncbi.nlm.nih.gov) [25] being the most widely used form of computational function prediction methods, assigning unannotated proteins with the function of their annotated inferred homologs [10]. However, this analysis is directly dependent on databases and the annotation observed for the retrieved sequences. For that reason, when transferring function from homology inference, it is important to consider that databases contain errors, caused mainly by automatic propagation of annotation errors transferred by homology [8] and this method is, perhaps, the most sensitive to these errors. Further, the resulted database sequences, although significantly similar to query sequence, may not represent a true homolog, or may represent a paralog, instead of an ortholog, or, further, even if an ortholog was retrieved, could not present the same function (Fig. 1b). Certainly, the expansion of databases of biological sequences brought another level of problem for functional assignment. Currently, most database sequences resulting from a similarity search are hypothetical proteins with unknown function, making the analysis unfruitful and frustrating or hiding more distant-related sequences containing reliable annotation. In general, the inference of function is reliable only for very high levels of sequence identity (roughly more than 60%) [26]. An alternative to BLAST analysis is the HMMER web server (www.ebi.ac.uk/Tools/hmmer) [18] that implements protein sequence databases searches through alignments using HMM. It claims to return more correct distantly related proteins than BLAST, but HMMER search is limited to amino acid level.

Sequence similarity does not directly reflect phylogeny and may misrepresent the evolutionary structure of a phylogenetic tree [27]. As homology is an evolutionary concept, methods to infer protein function that use sequence similarity search tools (e.g., BLAST) against sequence databases should not be viewed as "homology-based," but are, instead, "similarity-based." On the other hand, the real "homology-based" methods are those exploiting phylogenetic information.

2.1.2 Protein Families and Domain Search

Domain search also include sequence similarity, but focuses on conserved motifs found in protein families. It takes into account the modular nature of the proteins and is putative more sensitive

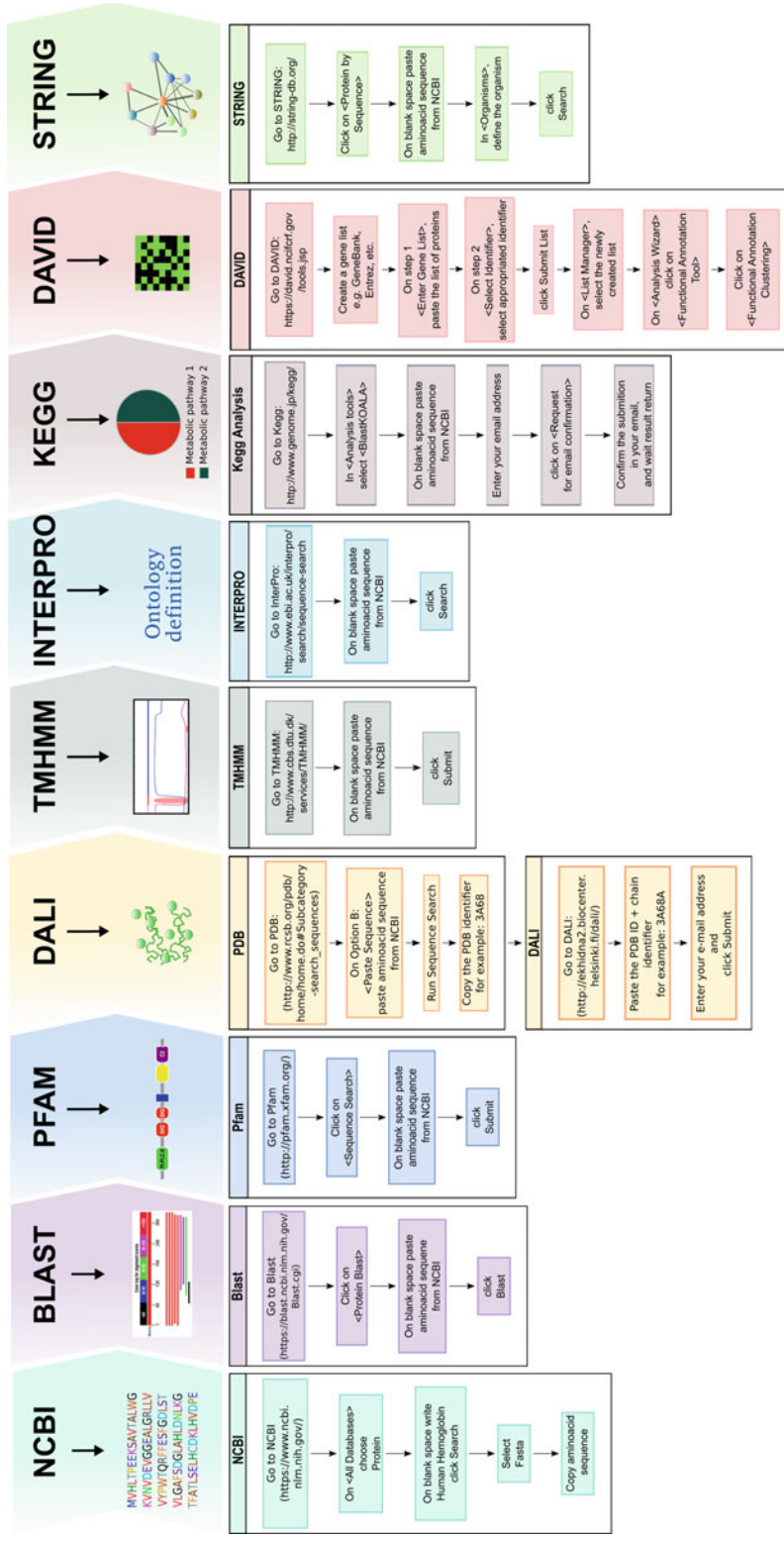


Fig. 3 Possible workflow to assign protein function using knowledge bases

because it considers only conserved regions, allowing detection of more distantly related proteins. The way used to establish motifs/domains in a protein family varies among different sources, but all start from multiple sequence alignments (MSA) of related (homologous) protein sequences in a given family. The conservation/variation in amino acids composition for each position in conserved functional regions (motifs/domains) are then extracted. The use of motifs/domains is tightly connected to protein families and can be extracted from MSA as separate single motifs/domains, multiple motifs/domains or even for the whole MSA. Conserved regions in motifs/domains are observed in MSA and described as: (a) *patterns*, a qualitative description of a motif/domain, indicating the occurrence of amino acids for each position of a motif/domain, represented through a regular expressions, as in the PROSITE database (prosite.expasy.org) [28]; (b) *profiles*, a quantitative description of a motif/domain, scoring the occurrence of each amino acid in MSA, as in Position-Specific Scoring Matrix (PSSM) used in the NCBI Conserved Domain Database (CDD) (www.ncbi.nlm.nih.gov/cdd) [29], or generating a probabilistic model using Hidden Markov Model (HMM) as in the Protein Family (Pfam) database (pfam.xfam.org) [18]; (c) *fingerprints*, groups of conserved and interrelated motifs capable to provide a signature for a particular protein family, as in the PRINTS database (www.bioinf.man.ac.uk/dbbrowser/PRINTS) [30]. These resources may be used in complementary to similarity search database analysis.

2.2 Structure-Based Methods

The function of a protein is inherently linked to its structure [31] and proteins sharing similar functions often have similar folds, a result originated from a common ancestral protein [2], the same homology concept used when comparing amino acid or nucleotide sequences. Sometimes, however, the function of one or both homologous proteins may change in the course of evolution while their folds remain largely unchanged, so in these cases the same fold may give rise to two functions [2, 26].

Methods to predict function from structure can be viewed according to the level of protein structure and specificity at which they operate, and be roughly separated in global fold similarity search and local structure definition or active site characterization [2, 31]. It should be noted, however, that not always global fold similarity correlates with functional similarity; examples include the TIM barrel fold, ferredoxin fold, and Rossmann fold global folds that are known to perform varying functions [31]. Functional assignment in these cases can be confirmed by local conservation of the residues [31]. The function of certain types of proteins is affected by a small number of residues found in a localized region of the three-dimensional structure. In enzymes, for example, the enzyme's catalytic function will be performed by a small number

of catalytic residues located in the active site [2]. Often, the specific arrangement and conformation of the residues are crucial to the performance of the function and remain strongly conserved over evolutionary time, even as the remainder of the protein's sequence and structure undergoes major changes [2]. Although global fold similarity can be used in many cases to assign a degree of functional similarity, predictions of specific biochemical or enzymatic function can be more accurately obtained from local fold similarity, i.e., in and around the protein active site [31].

Below the level of the fold come various other aspects of a protein's three-dimensional structure that may be associated with specific functions [2]. The surface of the protein, particularly its clefts and pockets, can hold important clues to function [2].

Many bioinformatics tools are available for structural function prediction. A hierarchical classification, including clusterization in homologous families, based on protein structures available in the Protein Data Bank (PDB) is presented by Class, Architecture, Topology and Homology (CATH) system (www.cathdb.info) [32] and Gene3D (gene3d.biochem.ucl.ac.uk) that uses information in CATH to predict the locations of structural domains on protein sequences from databases such as UniProtKB [33, 34]. Other methods exist for fold searching, including DALI (ekhidna.biocenter.helsinki.fi/dali_server) [35] and VAST (structure.ncbi.nlm.nih.gov/Structure/VAST) [36], which uses vector alignment of secondary structures, and CE (source.rcsb.org/jfatcatserver/ceHome.jsp) [37].

2.3 De Novo Protein Function Prediction

If an unknown protein has no significant similarity to any known protein, how is it possible to get insights about its function? In this case, computational approaches can be used to predict protein function de novo, that is, using only sequence or structure information to infer properties that are common to proteins of the same function [8]. These methods take the assumption that proteins of the same function are similarly adapted to same conditions (submitted to the same evolutionary constraints), such as pH, properties of a ligand, structural flexibility, etc. which will be reflected in their sequence and structural features [8]. Although not directly, these methods are also dependent on databases and proteins with already known function. This occurs because de novo methods generally use algorithms based on supervised learning models or statistical models, including Support Vector Machines (SVM), artificial neural networks, and Hiden Markov Model (HMM). These methods are usually less accurate than annotation transfer but are able to capture significant correlations between features and functions [8]. To do that, it needs to be "trained," that is, before scanning an amino acid sequence the models must be built from previously known proteins with the desired function or cellular

localization. These methods are largely used to establish functional residues or the subcellular localization of proteins [8].

Methods to predict functional residues assume that residues that have a similar function in different proteins are likely to possess similar physicochemical characteristics [8]. For example, residues that bind DNA share common structural and physicochemical features in most DNA-binding proteins (e.g., secondary structures, geometries, solvent accessibility, charge, hydrophobicity) [8]. There are several methods for the prediction of DNA- or metal-binding residues from sequence or structure [8].

Determining the subcellular localization of a protein helps to establish its function and can be very relevant for its experimental characterization [8]. Subcellular localization can also be predicted from similarity and motif searches if similar protein sequences with known function are available in databases, but de novo methods, instead, exploit the known correlation between amino acid composition and localization [8] and may help to even improve the knowledge about known proteins.

Many useful bioinformatics tools are available for online analysis; examples are: the Protein Subcellular Localization Prediction System (LocTree3; www.rostlab.org/services/loctree3) [38] that classifies proteins from eukaryotes, bacteria, and archaea; Balanced Subcellular Localization Predictor (BaCelLo; gpcr2.biocomp.unibo.it/bacello) [39], a predictor for the subcellular localization of proteins in eukaryotes; TargetP (www.cbs.dtu.dk/services/TargetP) [40], a predictor for eukaryotic proteins based on the presence of N-terminal signal peptide for chloroplast, mitochondrial, or secretory pathway; Subcellular Localisation Predictor (Protein Prowler; pprowler.imb.uq.edu.au) [41] determines the localization of the protein in secretory pathway, mitochondrion, or chloroplast; TMHMM (www.cbs.dtu.dk/services/TMHMM) [42] predicts transmembrane helices in protein sequences; LipoP (www.cbs.dtu.dk/services/LipoP) [43] predicts lipoproteins and signal peptides from Gram-negative bacteria protein sequences; TatP (www.cbs.dtu.dk/services/TatP) [44] predicts the presence and location of Twin-arginine signal peptide cleavage sites in bacteria.

2.4 Standard Vocabulary

Standard vocabulary on protein functional annotation provides important information to support researches on functional genomics, molecular and computational biology [4]. Schemes such as the enzyme classification system, or Enzyme Commission (EC), based on enzymatic reactions (www.chem.qmul.ac.uk/iubmb/enzyme) [45] that has been widely used in protein knowledge resources. Similarly, the Gene Ontology (GO) Consortium consists of standardized ontologies for describing gene function (www.geneontology.org) [46]. An ontology is a formal representation of knowledge by means of defined terms and its interrelationships,

allowing sequence annotation to different levels depending on the available information [46]. Both EC and GO are examples of frameworks that assign functions to groups of genes and gene products [47], creating controlled vocabulary and promoting database interoperability, but no system is directly based on protein sequences. More recently, a classification system was created for membrane transport proteins, named Transport Commission (TC), in analogy to EC system, based on the type of transport but in contrast to EC, also considers phylogenetic information based on families of homologous proteins involved (www.chem.qmul.ac.uk/iubmb/mtp) [48]. A number of other resources benefit from such controlled vocabulary, for example, the DAVID database (david.ncifcrf.gov) [49], which allows exploring functional annotation for large list of genes. EC, GO, and, more recently, TC numbers have been assigned to individual protein sequences in protein sequence databases such as UniProtKB, NCBI protein, and others. There are tools that combine standard vocabulary with similarity-based methods in predicting function from protein sequences, associating GO terms from similar proteins found in database, such as Gotcha [50] and PFP (kiharalab.org/web/pfp.php) [51], or combining different methods, including similarity and domain search, SVM and sequence derived protein features, such as CombFunc (www.sbg.bio.ic.ac.uk/~mwass/combfunc) [52] and ProtFun (www.cbs.dtu.dk/services/ProtFun) [53].

Different and complementary approaches have been applied for functional classification of proteins (and their genes) in large databases, mainly from predicted proteomes from complete genome sequences of all domains of life. These systems use bioinformatic algorithms and pipelines to generate clusters or families of protein sequences, assumed to be homologous, and classify them functionally. It is very useful in high-throughput analysis for functional classifications based on similarity search methods. Examples of those systems are The Clusters of Orthologous Groups (COG; www.ncbi.nlm.nih.gov/COG) [1], Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups (EggNOG; eggnogdb.embl.de) [54], Protein ANalysis THrough Evolutionary Relationships (PANTHER) Classification System (www.pantherdb.org) [55]. Other, special systems exist, dedicated to the classification of a more restricted group of function, for example, Carbohydrate-Active Enzymes (CAZy) database, dedicated to the families of enzymes that catalyze reactions (that degrade, modify, or create) glycosidic bonds (www.cazy.org) [56].

2.5 Systems Information

2.5.1 Genomic Context

In all organisms, the gene constitute a fundamental unit and its coded proteins tend to associate into higher levels of macromolecular complexes, biochemical pathways, and functional modules that are groups of interacting proteins acting together to accomplish a cellular process [16]. Now it is well recognized the

"modular nature" of cellular systems and this concept is considered a fundamental aspect of biological organization. Functional modules can be seen as a group of molecules acting in conjunction and interacting between them in order to perform a cellular/physiological function, with weaker connections to other functional modules [57, 58]. Frequently, functional modules show a high degree of conservation across species and may be identified in genomic associations (also linked to functional associations), such as conservation of gene order, gene/domain fusion events, and similarity of their phylogenetic profile [31, 59]. For example, the gene order is conserved in genes coding for enzymes or proteins involved in a particular metabolic pathways or cellular process, generally clustered in operons, and may serve as important clues for assigning functions if two genes retain close proximity even across large phylogenetic distances, indicating the presence of selective forces maintaining the gene organization [31]. Domain fusion is also another evolutionary event indicating functional associations in proteins, occurring when two functions are exerted by two independent proteins in one organism, but in a single protein, containing two domains in another one [31].

As an extension of genome context methods, a third indicative of functional association is the co-occurrence of genes, that is, the presence or absence of genes, known as phylogenetic profile, observed in genomes across different taxonomic groups [60]. The phylogenetic profile may be used to predict protein function by correlating the phylogenetic distribution of a query gene with that of known genes [31, 60]. The use of evolutionary information in the prediction of gene function is frequently referred as phylogenomics [61] and more elaborated methods infer function by building phylogenetic trees from homologs from known and unknown genes, generally presenting different functions assumed to rise from duplication events; the uncharacterized functions are then predicted by the phylogenetic positions relative to characterized genes [61]. Methods implemented in Orthostrapper and Function Through Evolutionary Relationships (SIFTER; sifter.berkeley.edu) [5] belong to this category.

This functional association may also be predicted via co-expression pattern in microarray analyses and/or mining literature [31]. Genome context can also be integrated with other levels of protein function information, as for example, standard vocabulary and network-based predictions. Some bioinformatics tools provide means to integrate all these levels of information, as for example, the KEGG pathway database [62] of metabolic pathway predicted from complete genome sequences, or the STRING database [63] of protein–protein interactions from different sources (including physical and functional evidences for association) and neighborhood, co-occurrence, and fusion for genes in genomic context.

2.5.2 Protein–Protein Interaction and Network-Based Prediction

One goal of modern biology is to group proteins into functional modules that act together to perform biological processes via direct and indirect interactions. The types of protein interaction within modules include physical interactions that generate protein complexes and biochemical associations [16]. Network-based predictions take advantage of these key features as gene products exhibit the tendency to associate into macromolecular complexes, biochemical pathways, and functional modules. Empirical observation shows that about 70–80% of interacting protein pairs share at least one function [24]. This observation is the rationale for methods to predict protein function using a network of protein–protein interaction, where proteins with unknown function can be assigned to the same function of known proteins interacting with them in a network. Protein–protein interaction networks can be reconstructed using proteomics, genomics, RNA expression (e.g., DNA microarrays, SGE, and RNA-seq) protein–protein interaction experiments (e.g., two-hybrid analysis, co-immunoprecipitation, and mass spectrometry), and bioinformatics approaches, which can reveal previously overlooked components and unanticipated functional associations [16, 64, 65]. The function of an unknown protein can be predicted based on its direct interactions, that is, its direct connections with known function of members observed in the network, or assisted by module, where first, groups of dense connections are identified in the network (modules), and then each module is separately annotated based on known functions of module members [66]. This approach assigns a function to an unclassified protein on the basis of function(s) present among the classified interacting proteins [24]. However, a disadvantage of this approach lies in the fact that, generally, there are few interactions observed between proteins with unknown and known functions [24].

The representation of protein–protein interactions as a network has the advantage to increase confidence levels for individual interactions and the possibility to uncover sets of protein–protein interactions that unexpectedly link diverse cellular processes or that indicate crosstalk between cellular compartments [65].

3 Final Remarks

As discussed in this chapter, the prediction of protein function is directly or indirectly dependent on proteins experimentally characterized, primary sequence and structure databases, and identification of homologous from direct sequence or structure comparison or extracted characteristics. Considering that experimentally characterized proteins are much fewer than uncharacterized proteins, and that the last continue to grow faster, automatic function prediction is the only suitable way to assign function to these "new" proteins. However, although much of these proteins with unknown

function may present homologous proteins with known function, a significant part represent orphan genes/proteins or are part of orthologous groups of unknown proteins. Further, even for unknown proteins that's function can be determined automatically, there are many reasons that makes this a complex task [3]: protein function can be studied from its molecular role to its metabolic or phenotypic effect in the whole cell; the experimental characterization of a protein is performed at a particular condition of temperature, pH, ligands concentration, etc., frequently given just partial description of its function; proteins are often multifunctional (Molecular Function and Biological Process ontologies have 30% and 60% of proteins in Swiss-Prot with more than one leaf term, respectively); annotation errors may occur due to experiment interpretation; and protein function is generally associated to gene names, difficult to predict in diverse isoforms.

Comparison of the accuracy (percentage in brackets) in predicting molecular function for experimentally characterized proteins, showed high variability in software using similarity-based methods: BLAST (75%), GeneQuiz (64%), and Gotcha (89%); and phylogeny-based methods: SIFTER (96%) and Orthostrapper (11%) [67]. A globally miss rate over 50% was found comparing the performance of Blast2GO, InterProScan, PANTHER, Pfam, and ScanProsite [68]. These results suggest the need to combine different methods when trying to predict protein functions. In a more complete survey, the performance of 54 methods for protein function prediction was evaluated by Radivojac et al. [3]. The authors established a cutoff of 60% amino acid sequence identity between an unknown and an experimentally annotated protein to be considered easy to annotate and determined its function and also observed that the overall accuracy in determining the Molecular Functional category is higher on single-domain proteins, compared to multidomain proteins [3]. The value of, at least, 60% sequence identity, and more likely closer to 80%, was also observed as required for the accurate transfer of the third level of EC classification [4].

When performing function prediction analysis important considerations should be taken into account, as outlined by Radivojac et al. [3]: (a) overall, BLAST seems ineffective at predicting functional terms in Biological Process ontology, possibly due to multiple roles played by orthologs; (b) studies have shown that correlation between sequence and function similarity is weak when applied to pairs of proteins and that domain assignments alone are not sufficient to resolve function; (c) for Molecular Function category, function prediction performance is accurate, but for Biological Process, the performance is worst; (d) methods that perform better integrate a variety of experimental evidence and weight different data appropriately for ontology terms.

A number of bioinformatics tools are available for protein function prediction and many of these tools were presented along

the text using the different methods described in this chapter. Many other useful tools are available and can be found listed and classified in reviews such as Watson et al. [2], Hawkins and Kihara [31], Friedberg [12], and Punta and Ofran ([8]—Supporting information).

References

1. Galperin MY, Makarova KS, Wolf YI, Koonin EV (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res 43: D261–D269. doi:10.1093/nar/gku1223
2. Watson JD, Laskowski RA, Thornton JM (2005) Predicting protein function from sequence and structural data. Curr Opin Struct Biol 15:275–284. doi:10.1016/j.sbi.2005.04.003
3. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML, Piovesan D, Casadio R, Wang Z, Cheng J, Fang H, Gough J, Koskinen P, Törönen P, Nokso-Koivisto J, Holm L, Cozzetto D, Buchan DWA, Bryson K, Jones DT, Limaye B, Inamdar H, Datta A, Manjari SK, Joshi R, Chitale M, Kihara D, Lisewski AM, Erdin S, Venner E, Lichtarge O, Rentzsch R, Yang H, Romero AE, Bhat P, Paccanaro A, Hamp T, Kaßner R, Seemayer S, Vicedo E, Schaefer C, Achten D, Auer F, Boehm A, Braun T, Hecht M, Heron M, Hönigschmid P, Hopf TA, Kaufmann S, Kiening M, Krompass D, Landerer C, Mahlich Y, Roos M, Björne J, Salakoski T, Wong A, Shatkay H, Gatzmann F, Sommer I, Wass MN, Sternberg MJE, Škunca N, Supek F, Bošnjak M, Panov P, Džeroski S, Šmuc T, Kourmpetis YAI, van Dijk ADJ, ter Braak CJF, Zhou Y, Gong Q, Dong X, Tian W, Falda M, Fontana P, Lavezzo E, Di Camillo B, Toppo S, Lan L, Djuric N, Guo Y, Vucetic S, Bairoch A, Linial M, Babbitt PC, Brenner SE, Orengo C, Rost B, Mooney SD, Friedberg I (2013) A large-scale evaluation of computational protein function prediction. Nat Methods 10:221–227. doi:10.1038/nmeth.2340
4. Clark WT, Radivojac P (2011) Analysis of protein function and its prediction from amino acid sequence. Proteins Struct Funct Bioinforma 79:2086–2096. doi:10.1002/prot.23029
5. Sahraeian SM, Luo KR, Brenner SE (2015) SIFTER search: a web server for accurate phylogeny-based protein function prediction. Nucleic Acids Res 43:W141–W147. doi:10.1093/nar/gkv461
6. Galperin MY, Koonin EV (2010) From complete genome sequence to "complete" understanding? Trends Biotechnol 28:398–406. doi:10.1016/j.tibtech.2010.05.006
7. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y (1998) Predicting function: from genes to genomes and back. J Mol Biol 283:707–725
8. Punta M, Ofran Y (2008) The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. PLoS Comput Biol 4: e1000160
9. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12:85–94
10. Sleator RD (2012) Prediction of protein functions. In: Kaufmann M, Klinger C (eds) Functional genomics. Springer, New York, NY, pp 15–24
11. Sleator RD, Walsh P (2010) An overview of in silico protein function prediction. Arch Microbiol 192:151–155. doi:10.1007/s00203-010-0549-9
12. Friedberg I (2006) Automated protein function prediction – the genomic challenge. Brief Bioinform 7:225–242. doi:10.1093/bib/bbl004
13. Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 8:995–1005. doi:10.1038/nrm2281
14. Khan I, Chen Y, Dong T, Hong X, Takeuchi R, Mori H, Kihara D (2014) Genome-scale identification and characterization of moonlighting proteins. Biol Direct. doi:10.1186/s13062-014-0030-9
15. Jeffery CJ (1999) Moonlighting proteins. Trends Biochem Sci 24:8–11
16. Hu P, Janga SC, Babu M, Díaz-Mejía JJ, Butland G, Yang W, Pogoutse O, Guo X, Phanse S, Wong P, Chandran S, Christopoulos C, Nazarians-Armavil A, Nasseri NK, Musso G, Ali M, Nazemof N, Eroukova V, Golshani A, Paccanaro A, Greenblatt JF, Moreno-Hagelsieb G, Emili A (2009) Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. PLoS Biol 7:e1000096. doi:10.1371/journal.pbio.1000096

17. Madupu R, Richter A, Dodson RJ, Brinkac L, Harkins D, Durkin S, Shrivastava S, Sutton G, Haft D (2012) CharProtDB: a database of experimentally characterized protein annotations. Nucleic Acids Res 40:D237–D241. doi:10.1093/nar/gkr1133
18. Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR (2015) HMMER web server: 2015 update. Nucleic Acids Res 43:W30–W38. doi:10.1093/nar/gkv397
19. Goodacre NF, Gerloff DL, Uetz P (2014) Protein domains of unknown function are essential in bacteria. mBio 5:e00744-13. doi:10.1128/mBio.00744-13
20. Bateman A, Coggill P, Finn RD (2010) DUFs: families in search of function. Acta Crystallogr Sect F Struct Biol Cryst Commun 66:1148–1152. doi:10.1107/S1744309110001685
21. Theißen G (2002) Orthology: secret life of genes. Nature 415:741–741. doi:10.1038/415741a
22. Zakon HH (2002) Convergent evolution on the molecular level. Brain Behav Evol 59:250–261
23. Doolittle RF (1994) Convergent evolution: the need to be explicit. Trends Biochem Sci 19:15–18. doi:10.1016/0968-0004(94)90167-8
24. Vazquez A, Flammini A, Maritan A, Vespignani A (2003) Global protein function prediction from protein–protein interaction networks. Nat Biotechnol 21:697–700
25. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. doi:10.1186/1471-2105-10-421
26. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y (2003) Automatic prediction of protein function. Cell Mol Life Sci CMLS 60:2637–2650. doi:10.1007/s00018-003-3114-8
27. Engelhardt BE, Jordan MI, Srouji JR, Brenner SE (2011) Genome-scale phylogenetic function annotation of large and diverse protein families. Genome Res 21:1969–1980. doi:10.1101/gr.104687.109
28. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I (2013) New and continuing developments at PROSITE. Nucleic Acids Res 41:D344–D347. doi:10.1093/nar/gks1067
29. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH (2015) CDD: NCBI's conserved domain database. Nucleic Acids Res 43:D222–D226. doi:10.1093/nar/gku1221
30. Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, Popov I, Roma-Mateo C, Theodosiou A, Mitchell AL (2012) The PRINTS database: a fine-grained protein sequence annotation and analysis resource – its status in 2012. Database 2012:bas019. doi:10.1093/database/bas019
31. Hawkins T, Kihara D (2007) Function prediction of uncharacterized proteins. J Bioinforma Comput Biol 5:1–30
32. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees JG, Lehtinen S, Studer RA, Thornton J, Orengo CA (2015) CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res 43:D376–D381. doi:10.1093/nar/gku947
33. Lam SD, Dawson NL, Das S, Sillitoe I, Ashford P, Lee D, Lehtinen S, Orengo CA, Lees JG (2016) Gene3D: expanding the utility of domain assignments. Nucleic Acids Res 44:D404–D409. doi:10.1093/nar/gkv1231
34. Yeats C, Lees J, Carter P, Sillitoe I, Orengo C (2011) The Gene3D web services: a platform for identifying, annotating and comparing structural domains in protein sequences. Nucleic Acids Res 39:W546–W550. doi:10.1093/nar/gkr438
35. Holm L, Rosenstrom P (2010) Dali server: conservation mapping in 3D. Nucleic Acids Res 38:W545–W549. doi:10.1093/nar/gkq366
36. Gibrat JF, Madej T, Bryant SH (1996) Surprising similarities in structure comparison. Curr Opin Struct Biol 6:377–385
37. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11:739–747
38. Goldberg T, Hecht M, Hamp T, Karl T, Yachdav G, Ahmed N, Altermann U, Angerer P, Ansorge S, Balasz K, Bernhofer M, Betz A, Cizmadija L, Do KT, Gerke J, Greil R, Joerdens V, Hastreiter M, Hembach K, Herzog M, Kalemanov M, Kluge M, Meier A, Nasir H, Neumaier U, Prade V, Reeb J, Sorokoumov A, Troshani I, Vorberg S, Waldraff S, Zierer J, Nielsen H, Rost B (2014) LocTree3 prediction of localization. Nucleic Acids Res 42:W350–W355. doi:10.1093/nar/gku396
39. Pierleoni A, Martelli PL, Fariselli P, Casadio R (2006) BaCelLo: a balanced subcellular localization predictor. Bioinformatics 22:e408–e416. doi:10.1093/bioinformatics/btl222
40. Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular

localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300:1005–1016. doi:10.1006/jmbi.2000.3903

41. Boden M, Hawkins J (2005) Prediction of subcellular localization using sequence-biased recurrent networks. Bioinformatics 21:2279–2286. doi:10.1093/bioinformatics/bti372
42. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes 11 Edited by F. Cohen. J Mol Biol 305:567–580. doi:10.1006/jmbi.2000.4315
43. Juncker AS, Willenbrock H, von Heijne G, Brunak S, Nielsen H, Krogh A (2003) Prediction of lipoprotein signal peptides in gram-negative bacteria. Protein Sci 12:1652–1662. doi:10.1110/ps.0303703
44. Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S (2005) Prediction of twin-arginine signal peptides. BMC Bioinformatics 6:167
45. du Plessis L, Skunca N, Dessimoz C (2011) The what, where, how and why of gene ontology – a primer for bioinformaticians. Brief Bioinform 12:723–735. doi:10.1093/bib/bbr002
46. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29
47. Lesk AM (2010) Introduction to protein science: architecture, function, and genomics, 2nd edn. Oxford University Press, Oxford
48. Saier MH (2006) TCDB: the transporter classification database for membrane transport protein analyses and information. Nucleic Acids Res 34:D181–D186. doi:10.1093/nar/gkj001
49. Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57. doi:10.1038/nprot.2008.211
50. Martin DM, Berriman M, Barton GJ (2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 5:178. doi:10.1186/1471-2105-5-178
51. Hawkins T, Luban S, Kihara D (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 15:1550–1556. doi:10.1110/ps.062153506
52. Wass MN, Barton G, Sternberg MJE (2012) CombFunc: predicting protein function using heterogeneous data sources. Nucleic Acids Res 40:W466–W470. doi:10.1093/nar/gks489
53. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Stærfeldt HH, Rapacki K, Workman C, Andersen CAF, Knudsen S, Krogh A, Valencia A, Brunak S (2002) Prediction of human protein function from post-translational modifications and localization features. J Mol Biol 319:1257–1265. doi:10.1016/S0022-2836(02)00379-0
54. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286–D293. doi:10.1093/nar/gkv1248
55. Mi H (2004) The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res 33:D284–D288. doi:10.1093/nar/gki078
56. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42:D490–D495. doi:10.1093/nar/gkt1178
57. Wagner GP, Pavlicev M, Cheverud JM (2007) The road to modularity. Nat Rev Genet 8:921–931. doi:10.1038/nrg2267
58. Pereira-Leal JB, Levy ED, Teichmann SA (2006) The origins and evolution of functional modules: lessons from protein complexes. Philos Trans R Soc B Biol Sci 361:507–517. doi:10.1098/rstb.2005.1807
59. Osterman A, Overbeek R (2003) Missing genes in metabolic pathways: a comparative genomics approach. Curr Opin Chem Biol 7:238–251. doi:10.1016/S1367-5931(03)00027-9
60. Kensche PR, van Noort V, Dutilh BE, Huynen MA (2008) Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. J R Soc Interface 5:151–170. doi:10.1098/rsif.2007.1047
61. Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 8:163–167. doi:10.1101/gr.8.3.163
62. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457–D462. doi:10.1093/nar/gkv1070
63. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic

M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43:D447–D452. doi:10.1093/nar/gku1003

64. Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast 18:523–531
65. Mayer ML, Hieter P (2000) Protein networks—built by association. Nat Biotechnol 18:1242–1243. doi:10.1038/82342
66. Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Mol Syst Biol. doi:10.1038/msb4100129
67. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE (2005) Protein molecular function prediction by Bayesian Phylogenomics. PLoS Comput Biol 1:e45. doi:10.1371/journal.pcbi.0010045
68. Rodrigues BN, Steffens MBR, Raittz RT, Santos-Weiss ICR, Marchaukoski JN (2015) Quantitative assessment of protein function prediction programs. Genet Mol Res 14:17555–17566. doi:10.4238/2015.December.21.28

Part II

DNA Analysis

Chapter 6

Capturing Three-Dimensional Genome Organization in Individual Cells by Single-Cell Hi-C

Takashi Nagano, Steven W. Wingett, and Peter Fraser

Abstract

Hi-C is a powerful method to investigate genome-wide, higher-order chromatin and chromosome conformations averaged from a population of cells. To expand the potential of Hi-C for single-cell analysis, we developed single-cell Hi-C. Similar to the existing "ensemble" Hi-C method, single-cell Hi-C detects proximity-dependent ligation events between cross-linked and restriction-digested chromatin fragments in cells. A major difference between the single-cell Hi-C and ensemble Hi-C protocol is that the proximity-dependent ligation is carried out in the nucleus. This allows the isolation of individual cells in which nearly the entire Hi-C procedure has been carried out, enabling the production of a Hi-C library and data from individual cells. With this new method, we studied genome conformations and found evidence for conserved topological domain organization from cell to cell, but highly variable interdomain contacts and chromosome folding genome wide. In addition, we found that the single-cell Hi-C protocol provided cleaner results with less technical noise suggesting it could be used to improve the ensemble Hi-C technique.

Key words Hi-C, Chromosome conformation capture (3C), Single-cell analysis, Chromatin interactions, Genome organization, In-nucleus ligation, In-solution ligation

1 Introduction

Mammalian chromatin is composed of approximately two meters of DNA and associated protein molecules, existing as a few dozen separate DNA threads known as chromosomes. Each chromosome carries a huge amount of genetic and epigenetic information which needs to be accessed in a highly organized manner, despite the fact it is folded within the nucleus as small as 5–20 μm in diameter. Decades of microscopic observation have shown that the chromatin is folded in nonrandom manner to form functionally relevant higher-order structures, but at the same time those structures vary from cell to cell [1]. While the inherent advantage of a microscopic approach is in obtaining information at the single-cell level, it is limited to small numbers of loci or regions precluding a comprehensive view of

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_6,

chromatin folding. Chromatin conformation capture (3C)-based methods including Hi-C provide alternative ways to address chromatin folding within the nucleus and complement the lack of throughput of microscopy approaches [2].

The principle of 3C-based methods is simple. In brief, the three-dimensional (3D) conformation of chromatin is preserved by formaldehyde cross-linking. Then the genomic DNA is digested with a restriction enzyme, followed by re-ligation. Re-ligation can occur between the adjoining or neighbor fragments on the primary DNA sequence, but also between fragments that are in close proximity in the 3D structure but not necessarily nearby on the primary sequence (proximity-dependent ligation; Fig. 1). All 3C-based methods work by detecting differences in the frequency of such proximity-dependent ligations, either over distance on the primary DNA sequence or between different chromosomes using various methods. For example, the original 3C method uses PCR to detect

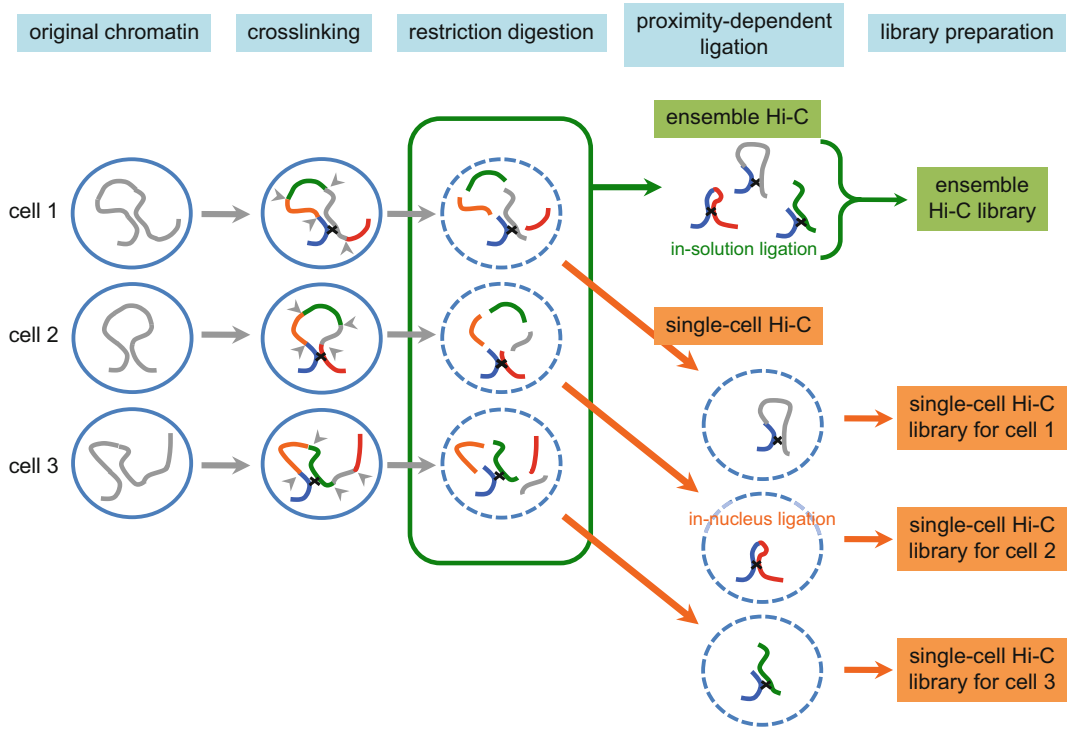


Fig. 1 Overview of ensemble Hi-C and single-cell Hi-C protocols. Ensemble Hi-C (*gray* and *green arrows*) and single-cell Hi-C (*gray* and *orange arrows*) share the same basic workflow, in which cross-linked chromatin is digested with a restriction enzyme. However, they differ at the proximity-dependent ligation step. Note that the chromatin conformation from each cell results in distinct patterns of proximity-dependent ligation. *Blue circle*, cell nucleus; *colored curve* in the circle, chromatin; *dotted circle*, permeabilized nucleus; *arrowhead*, restriction digestion site; ×, formaldehyde cross-link. (Modified with permission from Fig. 2a in Nagano T (2015) Higher-order chromatin organization revealed by single-cell Hi-C. Cell Technology 34:264–270 [Japanese])

such ligations, requiring knowledge on the loci of interest in advance. Hi-C on the other hand enables non-biased genome-wide detection of loci involved in the proximity-dependent ligations by labeling the ligation junctions with biotin to enrich and analyze the DNA around the junctions by the next-generation sequencing. The comprehensive nature of Hi-C data has permitted the elucidation of several important principles in higher-order chromatin structure such as compartmentalization within the nucleus [3] and organization of the genome in topologically associated domains (TADs) [4]. However, Hi-C and other 3C-based methods have an inherent weakness in that they assess ligation junctions from millions of cells together. This means no two ligation junctions can be assumed to be from the same cell or same chromosome. The analyses therefore represent the sum total of possible interactions for a given fragment. Due to the cell-to-cell variability in chromatin structure observed by microscopy, this information can only be used to approximate an average conformation.

To overcome this weakness, we developed the single-cell Hi-C method by modifying the original Hi-C protocol [5]. Essentially, the single-cell Hi-C differs from the original ensemble Hi-C in one point—the ligation step in single-cell Hi-C is performed in preserved nuclei (in-nucleus ligation) rather than a highly diluted solution of chromatin complexes (in-solution ligation) (Fig. 1). Our initial study using single-cell Hi-C [5] yielded 10 single-cell Hi-C datasets with sufficient coverage and quality to warrant further in-depth analyses among approximately 70 libraries prepared. We found that the chromosomes of each cell among the "homogeneous" cell population had a distinct 3D structure at the larger scale (for example, how each TAD is positioned in the chromosome territory), while all individual cells appeared to share the same TAD boundaries on the chromosomes. Increasing cell throughput to create a larger number of high-quality, high-coverage single-cell datasets will of course provide greater statistical power for discovery. Therefore this single-cell Hi-C protocol is still under development, and improvements are to be expected in the future.

The distinct nature of the single-cell Hi-C data compared to ensemble Hi-C allowed us to detect and remove some of the raw data that most likely represents technical noise. For example, single-cell Hi-C libraries are much less complex than ensemble Hi-C libraries, allowing us to reach saturation (sequence depth ~100×) with a relatively small number of read-pairs. In this situation, genuine read-pairs are expected to have similar levels of duplication in the raw sequence data. However, we found a substantial number of unduplicated read-pairs. In addition, each diploid cell has only two copies of DNA for autosomal genomic loci, meaning that a specific fragment end can only be ligated to two other ends in a single-cell dataset. In reality, however, we found fragment ends with more

than two ligations in many of our datasets. Importantly, when we removed the unduplicated read-pairs from the raw data we found fewer fragments with more than two ligations, suggesting they are in fact noise. These observations are not only useful to clean the single-cell Hi-C datasets but also suggest at least one source of experimental noise that arises in ensemble Hi-C as well, as ensemble and single-cell Hi-C share most of their experimental procedures. Therefore the information gained through single-cell Hi-C can be used to improve ensemble Hi-C, like the in-nucleus Hi-C ligation mentioned above [6].

2 Materials

2.1 Cell Fixation

1. Culture media and reagents to grow or maintain the cells of interest (the batch to use with formaldehyde should be at room temperature [20–25 °C]).
2. 16% Formaldehyde (methanol-free).
3. 2 M Glycine solution (can be stored at 4 °C for 1–2 weeks).
4. PBS: Phosphate-buffered saline, pH 7.4, Ca^{2+}-free, Mg^{2+}-free; chill on ice.
5. Liquid nitrogen (if you freeze down the fixed cells).

2.2 Hi-C Processing

1. Permeabilization buffer: 10 mM Tris–HCl pH 8.0, 10 mM NaCl, 0.2% (v/v) IGEPAL CA-630, cOmplete EDTA-free (Roche), prepare fresh and chill on ice.
2. 1.2× NEBuffer 3: Prepare fresh by diluting 10× NEBuffer 3 with nuclease-free water.
3. 20% SDS.
4. Temperature-controlled shaker for 1.5 mL tubes.
5. 20% Triton X-100 (prepare fresh).
6. Bgl II (50 U/μL).
7. 10 mM dCTP.
8. 10 mM dGTP.
9. 10 mM dTTP.
10. 0.4 mM biotin-14-dATP.
11. DNA polymerase I, large (Klenow) fragment (5 U/μL).
12. Nuclease-free water.
13. T4 DNA ligase reaction buffer (10×) (New England Biolabs).
14. 100× BSA (10 mg/mL).
15. T4 DNA ligase (1 U/μL) (Invitrogen).

2.3 Single-Cell Isolation

1. Cell strainer (30 or 40 μm).
2. Low-gelling temperature agarose.
3. PBS, pH 7.4 (Ca^{++}-free, Mg^{++}-free).
4. Phase contrast microscope.
5. Disposable Pasteur pipettes with mouth-controlled aspirator tube assembly.
6. Stereoscopic microscope.

2.4 Library Preparation from the Single-Cell Samples

1. Dynabeads M-280 Streptavidin.
2. Magnetic separation tube stand: for 1.5 mL tube, it is useful if you also have the stand for 0.2 mL PCR tubes.
3. Bead binding and washing buffer: 5 mM Tris–HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl.
4. 2× Bead binding buffer: 10 mM Tris–HCl pH 7.5, 1 mM EDTA, 2 M NaCl.
5. Rotating wheel.
6. Bead resuspension buffer: 10 mM Tris–HCl pH 7.5.
7. Nuclease-free water.
8. Alu I (10 U/μL; with reaction buffer).
9. Klenow fragment (3′ → 5′ exo-)(5 U/μL; with reaction buffer).
10. T4 DNA ligase (2000 U/μL; with reaction buffer) (New England Biolabs).
11. Oligonucleotide for library adapters (*see* Table 1).
12. Platinum Pfx DNA polymerase (with reaction buffer).
13. 10 mM each dNTP mixture.
14. Library amplification primers (*see* Table 1).
15. AMPure XP.
16. 70% Ethanol (prepare fresh).
17. 10 mM Tris–HCl pH 8.5.
18. Agarose (for gel electrophoresis).
19. Gel tank.
20. Orange G loading buffer: 10 mM Tris–HCl pH 7.5, 60 mM EDTA, 60% glycerol, 0.15% orange G.
21. 10 mg/mL ethidium bromide solution.
22. Transilluminator.
23. MinElute gel extraction kit (Qiagen).
24. 2100 Bioanalyzer system.
25. Illumina library quantification kit.

Table 1
Nucleotide sequences for indexed adapters and primers

[Oligonucleotides for CAA-indexed adapter].
(F) 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAA*T-3′
(R) 5′-pTTGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′
[Oligonucleotides for TAA-indexed adapter].
(F) 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTAA*T-3′
(R) 5′-pTTAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′
[Oligonucleotides for TCA-indexed adapter].
(F) 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCA*T-3′
(R) 5′-pTGAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′
[Oligonucleotides for ACC-indexed adapter].
(F) 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCTACC*T-3′
(R) 5′-pGGTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′
[Oligonucleotides for CCT-indexed adapter].
(F) 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCT*T-3′
(R) 5′-pAGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′
[Oligonucleotides for GTA-indexed adapter].
(F) 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTA*T-3′
(R) 5′-pTACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′
[Oligonucleotides for CAG-indexed adapter].
(F) 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAG*T-3′
(R) 5′-pCTGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′
[Oligonucleotides for TCG-indexed adapter].
(F) 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCG*T-3′
(R) 5′-pCGAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′
[Oligonucleotides for ATA-indexed adapter].
(F) 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCTATA*T-3′
(R) 5′-pTATAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′
[Oligonucleotides for TGC-indexed adapter].
(F) 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGC*T-3′
(R) 5′-pGCAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′
[Oligonucleotides for CTA-indexed adapter].
(F) 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTA*T-3′
(R) 5′-pTAGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′
[Oligonucleotides for GAG-indexed adapter].

(continued)

Table 1
(continued)

(F) 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAG*T-3′
(R) 5′-pCTCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′
[Library amplification primers].
(F) 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3′
(R) 5′-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T-3′

Sequence of indexed adapters and primers for the Illumina platform. The (p) and (*) denote phosphorylation and phosphorothioate bond, respectively. We recommend the HPLC purification grade. The oligonucleotides for adapters need to be annealed before use according to **Note 12**

26. Next-generation sequencing system (Illumina; note that the adapters and primers shown in this protocol are for Illumina platform).

3 Methods

3.1 Cell Fixation

1. Prepare the cells of interest (1–10 × 10^6) suspended in 21 mL of the medium appropriate for the cells at room temperature in 50 mL centrifuge tube.
2. Add 3 mL of 16% formaldehyde (final concentration 2%) and fix for exactly 10 min at room temperature by continuously and gently inverting the tube.
3. Add 1.632 mL of 2 M glycine (final concentration 0.127 M) to quench formaldehyde, mix well by inverting the tube several times, and incubate on ice for 5 min.
4. Spin the tube at 300 × *g* for 8 min at 4 °C.
5. Remove the supernatant (*see* **Note 1**).
6. Resuspend the cells with 50 mL of ice-cold PBS.
7. Spin the tube at 300 × *g* for 8 min at 4 °C.
8. Remove the supernatant (*see* **Note 1**).
9. If you don't immediately proceed to the Hi-C steps, you can snap freeze the fixed cell pellet with liquid nitrogen and store at −80 °C for several months.

3.2 Hi-C Processing

1. If you start from the frozen cell pellet, thaw on ice.
2. Resuspend the cell pellet with 50 mL of the permeabilization buffer (*see* **Note 2**) and incubate on ice for 30 min, with intermittent mixing by inverting the tubes for ~50 times every 5 min.
3. Spin the tube at 600 × *g* for 6 min at 4 °C.

4. Remove the supernatant leaving ~0.5 mL.
5. Resuspend the cells with the remaining supernatant and transfer to a new 1.5 mL tube.
6. Spin the tube at 600 × *g* for 6 min at 4 °C.
7. Remove the supernatant as much as possible without touching the cell pellet.
8. Add 800 μL of 1.2× NEBuffer 3 gently, not to disperse the cell pellet (*see* **Note 3**).
9. Spin the tube at 600 × *g* for 6 min at 4 °C.
10. Remove the supernatant as much as possible without touching the cell pellet.
11. Add 400 μL of 1.2× NEBuffer 3 gently, not to disperse the cell pellet (*see* **Note 3**).
12. Spin the tube at 600 × *g* for 6 min at 4 °C and remove the supernatant.
13. Add 400 μL of 1.2× NEBuffer 3 gently, not to disperse the cell pellet.
14. Add 6 μL of 20% SDS and mix gently by pipetting, not to make bubbles as much as possible.
15. Incubate at 37 °C for 60 min with shaking at 950 rpm.
16. Add 40 μL of 20% Triton X-100 and mix gently by pipetting, not to make bubbles as much as possible.
17. Incubate at 37 °C for 60 min with shaking at 950 rpm.
18. Add 30 μL of Bgl II (50 U/μL) and mix gently by pipetting, not to make bubbles as much as possible.
19. Incubate at 37 °C overnight with shaking at 950 rpm.
20. Add 1.5 μL of 10 mM dCTP, 1.5 μL of 10 mM dGTP, 1.5 μL of 10 mM dTTP, 37.5 μL of 0.4 mM biotin-14-dATP, and 10 μL of 5 U/μL DNA polymerase I Klenow fragment, and mix gently by pipetting, not to make bubbles as much as possible.
21. Incubate at 37 °C for 1 h with intermittently (10 s in every 30 s) shaking at 700 rpm.
22. Spin the tube at 600 × *g* for 6 min at 4 °C.
23. Transfer supernatant to a new tube, leaving 50 μL (*see* **Note 4**).
24. Prepare DNA ligation mix by pipetting 830 μL of nuclease-free water, 100 μL of 10× T4 DNA ligase buffer, 10 μL of 100× BSA, 10 μL of 1 U/μL T4 DNA ligase. Then add to the sample from **step 23** above (50 μL of supernatant and cell pellet) and mix gently by pipetting, not to make bubbles as much as possible (*see* **Note 5**).

25. Incubate at 16 °C for more than 4 h (overnight is fine; no shaking).
26. The cell suspension can be kept at 4 °C for several days at this point.

3.3 Single-Cell Isolation

1. Spin the post-ligation sample at 600 × *g* for 6 min at 4 °C.
2. Resuspend the pellet in 1 mL of PBS and remove large cell cluster by passing through a cell strainer with 30–40 μm mesh (*see* **Note 6**).
3. Count the cells and prepare ~100 μL of the suspension with ~150 cells/μL.
4. Melt 8 mg of low-gelling-temperature agarose in 1 mL of PBS in 1.5 mL tube in water bath at 70 °C (final agarose concentration 0.8%), then keep at 37 °C.
5. Warm 20 μL of PBS at 37 °C in 1.5 mL tube, add 5 μL of the nuclei suspension from **step 3** above and 25 μL of the molten agarose from **step 4** above, vortex gently to mix, and keep at 37 °C.
6. Warm a Pasteur pipette at 37 °C, pick up the mixture prepared at **step 5** above and make multiple small (diameter ≤ 0.5 mm) droplets onto a glass slide kept at 37 °C (*see* **Note 7**).
7. Put the glass slide on ice to harden the droplets.
8. Immerse the glass slide with agarose droplets into PBS in a Petri dish.
9. Observe each droplet under the phase contrast microscope to check the droplets containing single cells inside.
10. Pick up the droplets with single cells under the stereoscopic microscope using a new Pasteur pipette, and transfer into a 1.5 mL tube containing 25 μL of PBS one by one (*see* **Note 8**).

3.4 Hi-C Library Preparation from the Single-Cell Samples

1. Make sure that the tubes having single cells are securely closed, and spin briefly.
2. Incubate at 65 °C for overnight to reverse the cross-linking.
3. Mix Dynabeads M-280 streptavidin stock to get homogeneous suspension, take appropriate volume (25 μL per sample), put on a magnet stand, wait for 1 min, and remove supernatant on the magnet.
4. Take the bead-containing tube from a magnet, add ~1 mL of bead binding and washing buffer (*see* **Note 9**), tap the suspension to mix, spin briefly, put on a magnet again, wait for 1 min, and remove supernatant on the magnet.
5. Repeat the bead washing (**step 4** above) twice more.

6. After the third wash, take the bead-containing tube from a magnet, add appropriate volume (25 μL per sample) of 2× bead binding buffer, and mix by gentle pipetting.
7. Add 25 μL of the bead suspension to the sample from **step 2** above (total 50 μL per sample) and mix well by gentle vortex.
8. Put the samples on a rotator and incubate at room temperature for 1 hr. with rotating at 2 rpm (*see* **Note 10**).
9. Spin the tube briefly, put on a magnet, wait for 1 min, and remove supernatant on the magnet.
10. Take the tube from a magnet, add 200 μL of bead binding and washing buffer, tap the suspension to mix, spin briefly, put on a magnet again, wait for 1 min, and remove supernatant on the magnet.
11. Repeat the bead washing (**step 10** above) twice more.
12. After the third wash, take the bead-containing tube from a magnet, add 200 μL of bead resuspension buffer, tap the suspension to mix, spin briefly, put on a magnet again, wait for 1 min, and remove supernatant on the magnet.
13. Add 200 μL of bead resuspension buffer, tap the suspension to mix, spin briefly, put on a magnet again, and leave for ≥1 min until the Alu I reaction mix (below) is ready.
14. Prepare Alu I reaction mix by pipetting 44 μL of nuclease-free water, 5 μL of 10× reaction buffer (which comes with the enzyme), 1 μL of 10 U/μL Alu I; total 50 μL per sample (*see* **Note 11**).
15. Remove supernatant in the tubes from **step 13** above on the magnet, take the tubes from a magnet, and resuspend the beads with 50 μL of Alu I reaction mix from **step 14** above by gentle pipetting.
16. Incubate the tubes at 37 °C for 1 h on a rotator at 2 rpm (*see* **Note 10**).
17. Spin the tube briefly, put on a magnet, wait for 1 min, and remove supernatant on the magnet.
18. Take the tube from a magnet, add 200 μL of bead binding and washing buffer, tap the suspension to mix, spin briefly, put on a magnet again, wait for 1 min, and remove supernatant on the magnet.
19. Repeat the bead washing (**step 18** above) twice more.
20. After the third wash, take the bead-containing tube from a magnet, add 200 μL of bead resuspension buffer, tap the suspension to mix, spin briefly, put on a magnet again, wait for 1 min, and remove supernatant on the magnet.

21. Add 200 μL of bead resuspension buffer, tap the suspension to mix, spin briefly, put on a magnet again, and leave for ≥1 min until the A-tailing reaction mix (below) is ready.
22. Prepare A-tailing reaction mix by pipetting 43 μL of nuclease-free water, 5 μL of 10× reaction buffer (which comes with the enzyme), 1 μL of 10 mM dATP, 1 μL of 5 U/μL Klenow fragment (3′ -> 5′ exo-); total 50 μL per sample (*see* **Note 11**).
23. Remove supernatant in the tubes from **step 21** above on the magnet, take the tubes from a magnet, and resuspend the beads with 50 μL of A-tailing reaction mix from **step 22** above by gentle pipetting.
24. Incubate the tubes at 37 °C for 30 min on a rotator at 2 rpm (*see* **Note 10**).
25. Spin the tube briefly, put on a magnet, wait for 1 min, and remove supernatant on the magnet.
26. Take the tube from a magnet, add 200 μL of bead binding and washing buffer, tap the suspension to mix, spin briefly, put on a magnet again, wait for 1 min, and remove supernatant on the magnet.
27. Repeat the bead washing (**step 26** above) twice more.
28. After the third wash, take the bead-containing tube from a magnet, add 200 μL of bead resuspension buffer, tap the suspension to mix, spin briefly, put on a magnet again, wait for 1 min, and remove supernatant on the magnet.
29. Add 200 μL of bead resuspension buffer, tap the suspension to mix, spin briefly, put on a magnet again, and leave for ≥1 min until the adapter ligation reaction mix (below) is ready.
30. Prepare adapter ligation reaction mix by pipetting 41 μL of nuclease-free water, 5 μL of 10× reaction buffer (which comes with the enzyme), 2 μL of 2000 U/μL T4 DNA ligase; total 48 μL per sample (*see* **Note 11**).
31. Remove supernatant in the tubes from **step 29** above on the magnet, take the tubes from a magnet, and add 48 μL of adapter ligation reaction mix from **step 30** above.
32. Add 2 μL of 15 μM annealed indexed adapter (*see* **Note 12**) and mix the entire bead suspension well by gentle pipetting.
33. Incubate the tubes at room temperature for ≥30 min on a rotator at 2 rpm (*see* **Notes 10** and **13**).
34. Spin the tube briefly, put on a magnet, wait for 1 min, and remove supernatant on the magnet.
35. Take the tube from a magnet, add 200 μL of bead binding and washing buffer, tap the suspension to mix, spin briefly, put on a magnet again, wait for 1 min, and remove supernatant on the magnet.

36. Repeat the bead washing (**step 35** above) twice more.
37. After the third wash, take the bead-containing tube from a magnet, add 200 μL of bead resuspension buffer, tap the suspension to mix, spin briefly, put on a magnet again, wait for 1 min, and remove supernatant on the magnet.
38. Add 200 μL of bead resuspension buffer, tap the suspension to mix, spin briefly, put on a magnet again, and leave for ≥1 min until the PCR reaction mix (below) is ready.
39. Prepare PCR reaction mix by pipetting 36.2 μL of nuclease-free water, 5 μL of 10× amplification buffer (which comes with the enzyme), 2 μL of 50 mM $MgSO_4$ (which comes with the enzyme), 2 μL of 10 mM each dNTP mixture, 2 μL of 25 μM library amplification primer F, 2 μL of 25 μM library amplification primer R, 0.8 μL of Platinum Pfx DNA polymerase; total 50 μL per sample (*see* **Note 11**).
40. Remove supernatant in the tubes from **step 38** above on the magnet, take the tubes from a magnet, resuspend the beads with 50 μL of PCR reaction mix from **step 39** above by gentle pipetting, and transfer the suspension to PCR tubes.
41. Set the tubes to the thermal cycler and run the following program; 94 °C for 2 min, 25 cycles of [94 °C for 15 s, 62 °C for 30 s, 72 °C for 1 min], 72 °C for 10 min (*see* **Note 14**).
42. Tap the tubes to mix, spin briefly, put on a magnet, and transfer 45 μL of supernatant to new 1.5 mL tubes (*see* **Note 15**).
43. Add 81 μL of AMPure XP bead suspension to the 45 μL suspension from **step 42** (*see* **Note 16**), mix well by pipetting, and leave at room temperature for 5 min.
44. Spin briefly, put on a magnet for 3 min, and remove the supernatant.
45. Add 200 μL of freshly prepared 70% ethanol while keeping the tubes on a magnet, leave for ≥30 s, and remove supernatant.
46. Repeat the same wash as **step 45** above twice more.
47. After removing the supernatant of the last wash, remove the tubes from a magnet, spin briefly to collect the residual liquid at the bottom, put the tubes on a magnet again, wait for 1 min, and remove the residual liquid at the bottom as much as possible (*see* **Note 17**).
48. Open the tube lid on a magnet (but cover with clean paper towel, etc.) and wait until bead pellet dries (usually 5–15 min) (*see* **Note 18**).
49. Add 17 μL of 10 mM Tris–HCl pH 8.5 to the bead pellet on a magnet, then remove the tube from the magnet, resuspend the beads by pipetting, and incubate at room temperature for 5 min.

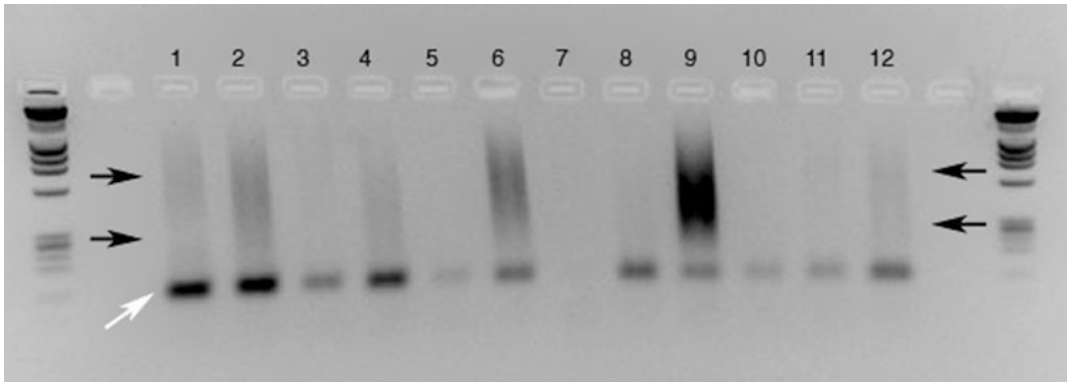


Fig. 2 An example gel electrophoresis image of the single-cell Hi-C libraries. Shown are 12 single-cell Hi-C libraries. Libraries are size-fractionated to collect fragments between 300–700 bp by extracting DNA from the gel piece of each lane between the two *black* arrows. The blobs shown by the *white arrow* are primer dimers, which should be removed during the size fractionation. (Reproduced with permission from Fig. 2 in Nagano T (2016) Single-cell Hi-C library construction to analyze the genome organization in individual cells. Experimental Medicine 34:1797–1806 [Japanese])

50. Spin briefly, put the tubes on a magnet for 3 min, and take the 15 μL of supernatant to a new tube.
51. Add 3 μL of orange G loading buffer to the 15 μL of supernatant from **step 50** above, mix well, load onto a 2% agarose gel, and run the sample until orange G migrate for about 3 cm (*see* **Note 19**).
52. Visualize DNA with ethidium bromide and observe the gel on the trans illuminator, which typically shows smears from approximately 300 bp to 1–2 kb (Fig. 2), and collect the gel piece corresponding to 300–700 bp (*see* **Note 20**).
53. Weigh the gel piece and collect DNA using Qiagen MinElute Gel Extraction Kit (*see* **Note 21**). The elute from the spin column is the final single-cell Hi-C library for sequencing if it satisfies the pre-sequencing check below (**step 54**).
54. Analyze the DNA size distribution by Agilent 2100 Bioanalyzer system (*see* **Note 22**).
55. When library size distribution is ok, quantify the library by qPCR.
56. Based on the quantification results, plan how much of each library is loaded in a lane (*see* **Note 23**).
57. Run the paired-end sequencing according to the manufacturer's instruction. This creates two FASTQ files (*see* **Note 24**).

3.5 Preliminary Data Analysis

1. The FASTQ files should be mapped independently (*see* **Note 25**) in single-end mode to obtain high-quality unique alignments (*see* **Note 26**).

2. The resulting aligned read files should be processed with scell_hicpipe (*see* **Note 27**), a software package tailored for analyzing single-cell Hi-C data (*see* **Note 28**). The scell_hicpipe scripts may be obtained from the Bitbucket Git hosting service: https://bitbucket.org/tanaylab/schic_pipeline.git using a web browser (*see* **Note 29**).
3. The Git repository downloaded from Bitbucket is compressed, but the files can be extracted on a Linux operating system with the command: "unzip [zip archive filename]."
4. The single-cell Hi-C analysis pipeline takes five files as input, which need to conform exactly to predefined formats, as described in the following steps (*see* **Note 30**). The first file should be generated by parsing the output from the aligner program to produce a list of paired-end reads. Each paired-end read should be written to a single line, listing in tab-separated format: Read1 chromosome name; Read 1 coordinate; Read 1 strand (+/−); Read 2 chromosome name; Read 2 coordinate; and Read 2 strand (+/−).
5. Create a second tab-delimited file summarizing the reference genome to which the FASTQ reads were aligned. The file should list chromosome names in the first column and the length (base pairs) of their respective chromosomes in the second column (*see* **Note 31**).
6. Generate a third tab-delimited file listing all the fragment ends (*see* **Note 32**). Each line should comprise an Index number (integer values starting 1 and incrementing by 1); Restriction fragment number; Strand (+/−); Chromosome name; Coordinate; Fragment length; and Fragment end length.
7. Make a fourth data file listing only the valid fragment ends. The file should be exactly the same format as that listing all the fragment ends (*see* **Note 33**).
8. Create a fifth data file comprising a one-column list of the index numbers (*see* **step 6** above in this section) of fragments ends classified as not valid.
9. Copy the first and second files (containing the single-cell Hi-C paired-end reads and the chromosome lengths) into the "input" folder (found within the downloaded archive). The remaining third, fourth, and fifth files (containing the fragment end data) should be copied into a newly made folder named "fends," which subsequently should also be copied to the input folder.
10. Running the pipeline requires a configuration file. The configuration file used to process the test dataset, which is found in the "input" folder and named "cell5.cfg," may be adapted to generate a new configuration file customized to the dataset to be processed.

11. The configuration file is used to specify the files that need processing and to override the default settings listed in the "makefile" (*see* **Note 34**). To view the makefile paramaters, type "cat makefile" on the command line. The options that may be overwritten will be found in the "Parameters" and "Output contact maps parameters" sections of the makefile. For example, the parameter "READ_PAIRS_FN?" is used to specify the file listing paired-end reads. Each makefile parameter has an adjacent comment, briefly describing its function.
12. Process the data and perform the quality control steps by moving to the folder containing the makefile and entering the following on the command line:

    make CFG = input/[name of configuration file].
13. While in the same directory, generate contact maps by entering the following on the command line: make plot_cmap CFG = input/[name of configuration file].

4 Notes

1. To remove the supernatant completely, spin the tube briefly after removing most of the supernatant to collect the remainder at the bottom and remove it with a fine pipette tip without touching the cell pellet.
2. First resuspend the cell pellet with small volume (~1 mL) of the permeabilization buffer and make sure the suspension is homogenous.
3. If you start with small number of cells (e.g., ~1×10^6) and disperse the cells at this point, they tend to attach the tube broadly after the next spin and may not form a clear pellet. If the cells are dispersed in 1.2× NEBuffer 3, you will be able to have the clear pellet by adding IGEPAL CA-630 at the final concentration of 0.02% and spin at 600 × *g* for 6 min at 4 °C.
4. To do this reproducibly, prepare the same tube with 50 μL of liquid and use this tube as a guide to show how 50 μL looks like in the tube.
5. In contrast to the original Hi-C, setting up the ligation without treating the samples with SDS after **step 21** makes the nuclei preserved during ligation. This enables the proximity-dependent ligation happen within the individual nuclei as "in-nucleus ligation."
6. When all the supernatant is replaced with PBS, some cells may stick to the tube and tip in the absence of detergent, but this may not be a big problem if you only plan to pick up tens of single cells in the downstream steps later.

7. Work quickly to make as many droplets as possible before the agarose suspension solidifies.
8. Do not reuse the Pasteur pipette to avoid cross-contamination between samples. The remaining cells after single-cell isolation can be used to extract ensemble Hi-C DNA for quality control purposes (restriction digestion efficiency, biotin labeling efficiency, etc.).
9. Use at least twice volume of the bead binding and washing buffer as the original bead suspension volume (split the beads into multiple tubes if necessary).
10. Make sure the beads don't settle in the tube (if they do, adjust the rotation speed, etc.).
11. Prepare some extra volume (for example, prepare for 13 samples when working with 12 samples) to equally cover all samples.
12. To sequence multiple single-cell libraries in the same lane, the indexes of the adapter should be sample-specific. In our previous study [5], we had the oligonucleotides listed in Table 1 from the commercial source as 100 μM solutions in nuclease-free water (two oligonucleotides are necessary for each indexed annealed adapter). To prepare the 15 μM stock of the annealed adapter, mix 15 μL each of F and R oligonucleotides for the same index with 70 μL of nuclease-free water in a PCR tube, and incubate by the thermal cycler at 95 °C for 5 min followed by 1 °C decrement per minute until 25 °C and keep at 25 °C for 30 min. The annealed 15 μM indexed adapter stocks can be stored at −20 °C as small (ideally for single use) aliquots. Note that these adapters are for Illumina platform.
13. The reaction can be extended to overnight if this is convenient.
14. The samples can be stored at −20 °C after the PCR.
15. First transfer the entire suspension to 1.5 mL tubes if you use the magnet for 1.5 mL tubes, but you can separate the supernatant from beads in the PCR tubes if you use the magnet for 0.2 mL PCR tubes. The supernatant will be more than 45 μL, but transfer 45 μL only to keep the sample volume in the following step constant and not to take the existing beads into the following steps.
16. Mix AMPure XP beads well to get a homogeneous suspension, take an aliquot, and equilibrate to room temperature before use.
17. Take care not to touch the bead pellet with pipette tip.
18. When the bead pellet is dry, it doesn't look glossy and has cracks.

19. Adding ethidium bromide to the gel and running buffer at the final concentration of 0.5 μg/mL helps to proceed to the following step quickly after the run.
20. The blob near 130 bp is primer dimer and shouldn't be collected. To protect DNA from degradation, quickly mark the gel to guide the collection and minimize UV irradiation.
21. Follow the manufacturer's protocol except for the following four points – (a) dissolving the gel piece in buffer QG is at room temperature with constant agitation for 30 min on a rotator; (b) to bind the DNA to the spin column, reapply the flow-through once more; (c) wash the spin column three times with buffer PE; (d) to elute DNA from spin column, repeat two independent elution with 10 μL of buffer EB each (total volume will be ~19 μL).
22. It is important to check if the library is contaminated with primer dimer or not (Fig. 3). The ratio of primer dimer in sequencing results is significantly more than the ratio at pre-sequencing stage, because primer dimer is more efficient to form clusters in flowcell compared to normal library molecules. When primer dimer occupies >5% of total by Bioanalyzer at this stage like Fig. 3c, it is worth considering to go back to **step 51** to re-run the sample on a gel and do the size-selection

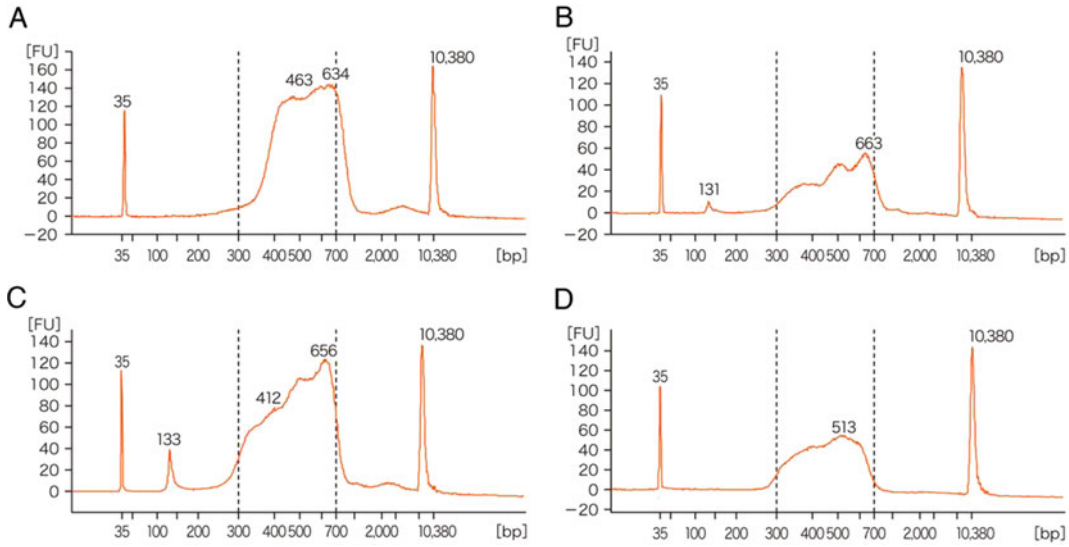


Fig. 3 Size distribution results of the example libraries by the Agilent 2100 Bioanalyzer system. The examples shown as A, B, and C derive from the lanes 9 (after 10× dilution), 11, and 12 in Fig. 2, respectively. The peaks near 130 bp in B and C are contamination by primer dimers. If the primer dimer contamination is obvious as in C, it is recommended to re-purify the library to reduce contamination. Panel D shows the re-purified library from the sample shown in C. Peaks at 35 bp and 10,380 bp are size markers (not part of the libraries). (Reproduced with permission from Fig. 2 in Nagano T (2016) Single-cell Hi-C library construction to analyze the genome organization in individual cells. Experimental Medicine 34:1797–1806 [Japanese])

again. But you will lose DNA during the re-purification, so this option is not realistic when DNA concentration is too low.

23. In our previous study using mouse diploid cells [5], we sequenced our single-cell Hi-C libraries under the condition where each dataset is expected to have an average of ≥2 million raw read-pairs (including duplication). This was done by loading up to 12 single-cell Hi-C libraries as an equimolar mixture in one lane of Illumina Genome Analyzer IIx, and we could sequence each library to nearly saturation with typical sequence depth of ~100×.
24. In our previous study using mouse diploid cells [5], we sequenced in 2 × 37–50 bp paired-end mode. When the indexed adapters in Table 1 are used, the index read is a part of each read (i.e., first three nucleotides; the first three nucleotides in each read of a read-pair should match).
25. Paired-end mapping using both FASTQ files should not be performed since this process usually assumes that the DNA being mapped forms a contiguous molecule when positioned on the reference genome. This will not be true for valid Hi-C di-tags.
26. We used the Maq aligner program for this step, using the default parameters and keeping the read-pairs in which both ends mapped uniquely with high-quality scores (MapQ ≥ 30) to the relevant reference genome.
27. Run the pipeline on a Linux operating system.
28. The single-cell Hi-C pipeline comprises three main steps, namely the processing of paired-end reads; the production of quality control metrics and charts; and finally the generation of contact maps.
29. Alternatively, on systems with Git installed, the full repository may be downloaded with the command: "git clone https://bitbucket.org/tanaylab/schic_pipeline.git". The Amos Tanay Group's website also hosts the software scripts: http://compgenomics.weizmann.ac.il/tanay/?page_id=580.
30. The user needs to create these data files which will be specific to the mapping results, Hi-C restriction enzyme used, and the chosen reference genome. Before analyzing real data, we suggest testing scell_hicpipe by processing a pre-made dataset in which the results are already known. The test dataset can also be downloaded from either Bitbucket or the Amos Tanay Group pages.
31. This information should be available from the source where the reference genome was obtained.
32. When following a Hi-C double-digest protocol in which a second restriction digestion step is performed instead of

sonication during library preparation, a fragment end is defined as the sequence between the first restriction enzyme cut site (mostly Bgl II in our previous study) and the second restriction enzyme cut site (Alu I in our previous study) [5]. These coordinates can be determined by performing an in silico digests of the reference genome FASTA files.

33. A restriction fragment may not always possess two valid fragment ends since not all sites cut by the first restriction enzyme are delimited by a pair of sites cut by the second restriction enzyme. Although choosing a six-cutter for the first restriction enzyme and a four-cutter for the second restriction enzyme will favor the production of both fragment ends, it is not guaranteed. Another reason to reject a fragment end is if it comprises a non-unique genomic sequence, and consequently reads should not align uniquely at this location using the mapping parameters described previously. Finally, the restriction fragment ends with significantly high coverage are liable to reflect errors in the reference genome and should be considered invalid [5].
34. A makefile is a special file listing commands to be executed. To run a makefile the user needs to navigate to the folder containing the makefile and then enter "make" on the command line.

References

1. Lanctôt C, Cheutin T, Cremer M et al (2007) Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. Nat Rev Genet 8:104–115
2. Dekker J, Marti-Renom MA, Mirny LA (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet 14:390–403
3. Lieberman-Aiden E, van Berkum NL, Williams L et al (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326:289–293
4. Dixon JR, Selvaraj S, Yue F et al (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485:376–380
5. Nagano T, Lubling Y, Stevens TJ et al (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502:59–64
6. Nagano T, Várnai C, Schoenfelder S et al (2015) Comparison of Hi-C results using in-solution versus in-nucleus ligation. Genome Biol 16:175

Chapter 7

Genome-Wide Cell Type-Specific Mapping of In Vivo Chromatin Protein Binding Using an FLP-Inducible DamID System in *Drosophila*

Alexey V. Pindyurin

Abstract

A thorough study of the genome-wide binding patterns of chromatin proteins is essential for understanding the regulatory mechanisms of genomic processes in eukaryotic nuclei, including DNA replication, transcription, and repair. The DNA adenine methyltransferase identification (DamID) method is a powerful tool to identify genomic binding sites of chromatin proteins. This method does not require fixation of cells and the use of specific antibodies, and has been used to generate genome-wide binding maps of more than a hundred different proteins in *Drosophila* tissue culture cells. Recent versions of inducible DamID allow performing cell type-specific profiling of chromatin proteins even in small samples of *Drosophila* tissues that contain heterogeneous cell types. Importantly, with these methods sorting of cells of interest or their nuclei is not necessary as genomic DNA isolated from the whole tissue can be used as an input. Here, I describe in detail an FLP-inducible DamID method, namely generation of suitable transgenic flies, activation of the Dam transgenes by the FLP recombinase, isolation of DNA from small amounts of dissected tissues, and subsequent identification of the DNA binding sites of the chromatin proteins.

Key words *Drosophila melanogaster*, Chromatin proteins, Genomic binding sites, DamID, Cell type-specific profiling, "Flp-Out" approach

1 Introduction

In eukaryotes, DNA replication, transcription, and repair are regulated by the chromatin structure [1–4]. Hence, a detailed understanding of the mechanisms underlying these processes requires precise information on the genome-wide binding patterns of chromatin components. Two methods are commonly used to define these patterns: chromatin immunoprecipitation (ChIP) and Dam identification (DamID) [5–7].

ChIP is based on reversible covalent cross-linking of protein–DNA complexes by chemical agents (such as formaldehyde) or UV light. After fixation, cells are lysed, the cross-linked chromatin is fragmented and specific antibodies are used to

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_7,

immunoprecipitate complexes containing a protein of interest (POI). Next, cross-link is reversed, POI-associated DNA fragments are purified, PCR-amplified and identified either by hybridization to a DNA microarray (ChIP-chip) or by high-throughput sequencing (ChIP-seq) [8–10]. Although ChIP can provide high-resolution mapping of protein binding sites [11], the availability of highly specific and effective antibodies is crucial for a successful experiment [12].

DamID does not require the use of antibodies, as this method relies on the specific activity of *E. coli* DNA adenine methyltransferase (Dam) [13, 14]. In vivo expression of the Dam-POI fusion proteins results in methylation at the N^6 position of adenine of genomic GATC sequences located near the Dam-POI binding sites (Fig. 1). Because endogenous adenine methylation is virtually absent in higher eukaryotes [15], the Dam-dependent methyladenine modifications are likely to have no effect on genome functioning. The methylation can be easily detected by using the DpnI restriction enzyme, which cuts only methylated GATC sequences. DpnI-digested genomic DNA is ligated to a specific adapter and then treated with the DpnII restriction enzyme, which cleaves the internal non-methylated GATC sequences. DNA fragments with adapters at both ends are amplified by PCR and sequenced.

By default, the resolution of DamID cannot be higher than the length of GATC fragments. In *Drosophila melanogaster*, there are 358,500 GATC fragments per haploid genome (release 5; excluding "U" and "Uextra" chromosome sequences) and between 50–75% of these fragments have sizes that allow successful PCR amplification (between 0.1 and 3 kb; with a median size of 200 bp). The range of amplified DNA fragments could be potentially adjusted by changing the concentration and/or sequence of the primer used for PCR [16].

To minimize non-targeted methylation, the Dam proteins should be expressed at a low level. In *Drosophila*, this goal is typically achieved exploiting the *heat shock protein 70* (*hsp70*) gene promoter; experiments are performed in the absence of heat shock or using a truncated version of the promoter that is not heat sensitive [13, 17]. In addition, to correct for background (non-targeted) Dam-dependent DNA methylation, Dam-POI profiling experiments should be performed in parallel with control experiment with cells expressing unfused Dam ("Dam only") [14].

For a correct interpretation of the DamID results, the properties of the system should be taken into account. The establishment of DNA methylation pattern by the Dam proteins requires their expression for several (usually 24) hours [7, 18]. Within its recognition site, Dam methylates the adenines on both DNA strands in a processive manner (i.e., almost simultaneously) [19, 20]. Then, during DNA replication, the GATC sequences methylated on both strands become hemimethylated and short time is required

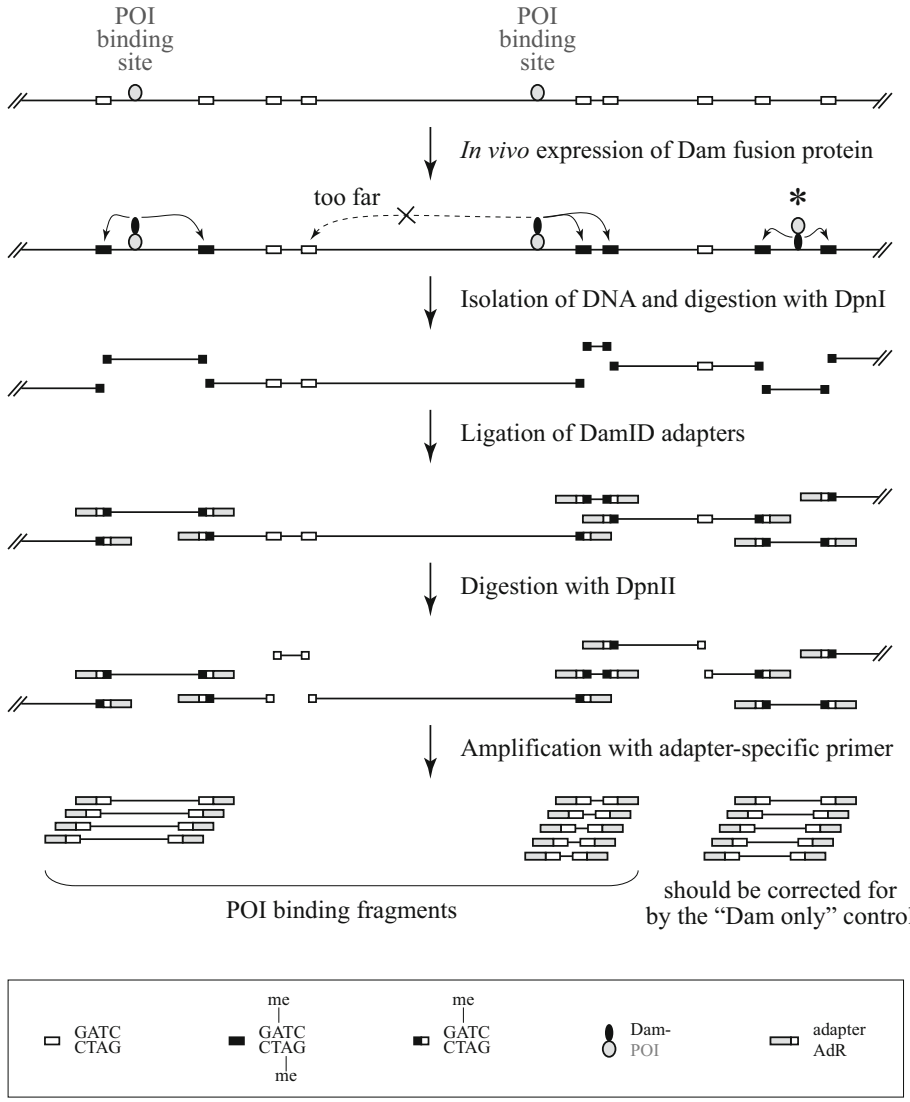


Fig. 1 Basic DamID analysis. Schematic representation of a short genomic region with POI binding sites. The Dam-POI fusion protein interacts with these sites and also with another ("background") site to which Dam has an affinity to bind (*asterisk*). Dam methylates adenines within GATC motifs located near the fusion protein binding sites. DNA fragments between two methylated GATC motifs are amplified for their subsequent identification as follows. Isolated genomic DNA is digested with the DpnI restriction enzyme, which cuts only methylated GATC sequences. Next, an adapter is ligated to the ends of the DpnI-cut fragments. Then, fragments with internal (non-methylated) GATC sequences are cleaved by the DpnII restriction enzyme. Finally, fragments with adapters at both ends are amplified by PCR. Control experiment with unfused Dam ("Dam only") performed in parallel allows correction for background (non-targeted) DNA methylation. me, methyl group

to restore the methylation of the second strand. Because the DpnI enzyme is much more specific for fully methylated than for hemi-methylated DNA [21], some recently replicated hemimethylated DNA fragments might not be detected. However, this is a relevant

issue only when a synchronized population of cells is used for the DamID experiments.

The DamID approach has been used to map the DNA binding sites of different chromatin proteins in a variety of organisms ranging from plants to mammals. The majority of experiments have been performed either in a whole organism (i.e., a heterogeneous mix of different tissues and cell types) or in cultured cell lines [17, 22–28]. However, the properties of DamID make it particularly suitable for cell type-specific profiling of chromatin proteins. Indeed, if the expression of a Dam protein is restricted to a particular tissue or cell type, then due to exquisite specificity and sensitivity of the DamID protocol [29, 30] the subsequent sorting of these cells or their nuclei [31, 32] is not required. This substantially simplifies the profiling procedure and minimizes possible artifacts associated with tissue disruption and sorting of specific types of cells or nuclei. Two different DamID modifications for cell type-specific profiling have been recently developed in *Drosophila*: the GAL4-inducible Targeted DamID (TaDa) procedure [33, 34] and the FLP-inducible STOP#1-Dam method [35]. Here, I describe in detail the latter approach.

The FLP-inducible STOP#1-Dam method is an adaptation of the "flp-out" approach [36] to DamID. Expression of Dam proteins, driven by the minimal *hsp70* promoter [37], is possible only after FLP-mediated excision of the transcriptional terminator, which is located between two directly oriented FRT sites in a DamID transgene (Fig. 2). The transcriptional terminator (STOP#1) sequence consists of yeast *His3* and the SV40 polyadenylation signal regions, a false translation initiation signal, and a 5′ splice donor site [38]; STOP#1 comparison with other transcriptional terminators showed that it is highly efficient in regulating the expression of the DamID transgene [35]. The cell type and tissue specificity of FLP-mediated DamID profiling largely depends on the specificity of the FLP protein expression. The higher the specificity of the recombinase expression, the lower the noise in the POI DamID binding profile. Many transgenic lines expressing FLP in specific subsets of cells already exist, including repo-FLP [39], dac-FLP [40], ey-FLP [41], ovo-FLP [42], GMR-FLP [43], R57C10-FLP [44], and a set of about 1000 enhancer-trap-FLPx2 lines [45]. If needed, a combination of a UAS-FLP and a suitable GAL4 driver can be used, although this requires several genetic crosses. Because FLP-inducible DamID construct does not contain a UAS element, profiling can be carried out in a GAL4-dependent mutant background (knockdown or overexpression of a specific gene). In addition, the DamID method can be particularly useful for studying the effects of point mutations (substitutions or deletions of amino acid residues) affecting the POI interaction with the chromatin, as

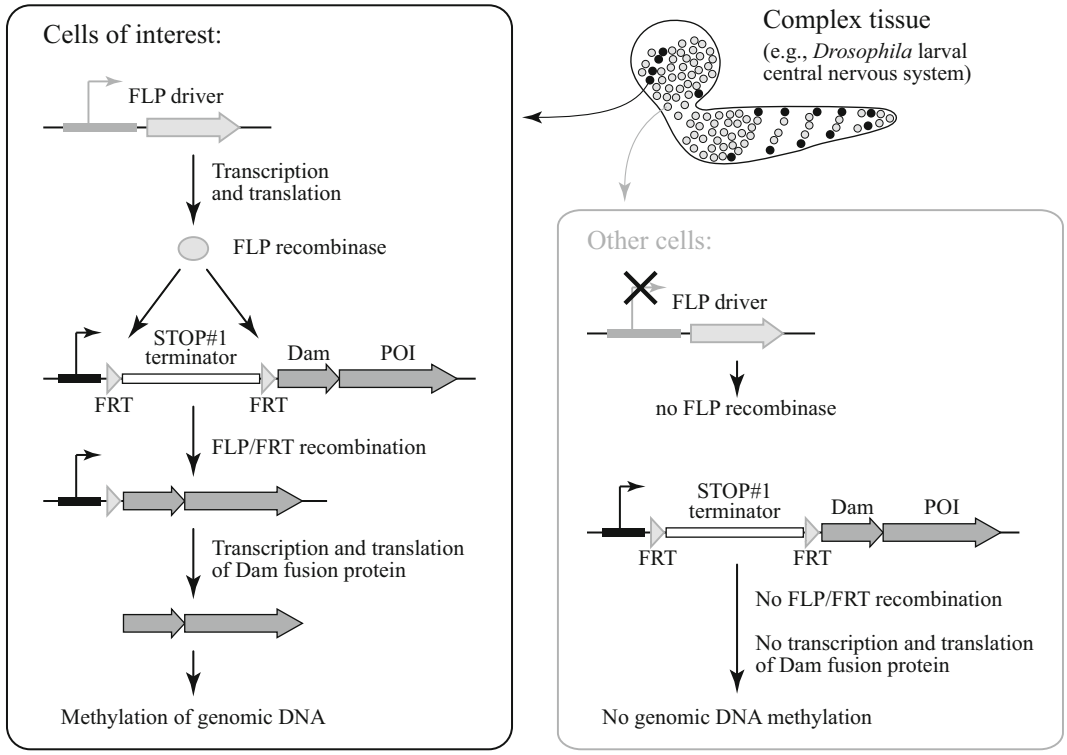


Fig. 2 The FLP-inducible DamID system. Expression of the FLP recombinase in a cell type of interest (which might constitute a minority of the cells in a heterogeneous tissue) using a cell-specific driver leads to the excision of the STOP#1 cassette flanked by the FRT sites from the DamID transgenic construct, allowing transcription of the Dam-POI coding gene driven by the ubiquitously active minimal *hsp70* promoter. This results in cell type-specific genomic DNA methylation by the Dam enzyme

there is no need of an antibody that recognizes the mutated protein [46, 47].

It is worth mentioning that a very small amount of dissected material is required for amplification of Dam methylated DNA fragments. The minimum number of cells (or the fraction of cells within a tissue) required for a reliable FLP-inducible DamID analysis is not known. However, because Dam methylated DNA fragments can be in principle amplified from a single cell [30], it is likely that the method could work even with a few Dam-POI expressing cells within a sample of heterogeneous tissue. For example, a single *Drosophila* brain from a third instar larva with a Dam-POI expressed in ~10% of the cells might be enough [35]. Finally, since the preparation of the DamID samples does not require expensive equipment and reagents (except for the step of high-throughput sequencing), the FLP-inducible DamID method can be afforded by most *Drosophila* laboratories.

2 Materials

2.1 Drosophila Equipment

For the full list of the minimal equipment required for working with flies, including transgenesis, *see* [48, 49].

1. Binocular microscope.
2. Fluorescent binocular microscope.
3. Drosophila anesthesia CO_2 station.
4. Injection apparatus and needles (optional, *see* below).
5. Forceps and dissection needles.
6. Pyrex® 9 depression glass spot plate.
7. Pellet pestles.
8. Equipment for working with *Drosophila* cell cultures (laminar hood, cell counter, incubator, etc.; optional, *see* below).

2.2 Molecular Biology Equipment

1. NanoDrop®ND-1000 spectrophotometer or equivalent (Thermo Scientific).
2. ABI PRISM® 3700 DNA Analyzer or equivalent (Applied Biosystems).
3. S2 ultrasonicator (Covaris) (optional, *see* below).
4. Magnetic particle concentrator (optional, *see* below).
5. 2100 Bioanalyzer instrument (Agilent) (optional, *see* below).
6. HiSeq 2000 instrument or equivalent (Illumina) (optional, *see* below).

2.3 Disposables

1. 1.5 mL centrifuge tubes.
2. 2.0 mL centrifuge tubes.
3. 0.2 mL PCR tubes.
4. Pipette tips (10, 200 and 1000 μL).
5. Pipette filter tips (10 and 200 μL).
6. Sterile 0.22 μm filters.
7. 21G needles and 1.0 mL syringes.
8. Microcon-30 kDa centrifugal filter units with Ultracel-30 membrane (Merck Millipore).
9. microTUBE AFA fiber pre-slit snap-cap 6 × 16 mm (Covaris) (optional, *see* below).

2.4 Reagents

1. Nuclease-free water.
2. Standard molecular biology reagents for molecular cloning (primers, high-fidelity DNA polymerase, restriction enzymes, etc.).

3. Gibson assembly® master mix (New England Biolabs) (optional, *see* below).
4. BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems).
5. Phosphate-buffered saline (PBS).
6. Ethanol (96%).
7. Isopropanol.
8. Agarose.
9. Ethidium bromide.
10. Phenol:chloroform:isoamyl alcohol (25:24:1 vol:vol:vol).
11. Chloroform.
12. Glycogen (20 mg/mL).
13. Proteinase K (20 mg/mL).
14. RNase A (100 mg/mL).
15. DNA mass ladder (0.1–10 kb).
16. SuRE/Cut Buffer H (10×) (Roche).
17. DpnI restriction enzyme (New England Biolabs; supplied with NEBuffer 4 (10×)).
18. DpnII restriction enzyme (New England Biolabs; supplied with NEBuffer DpnII (10×)).
19. T4 DNA ligase (Roche; supplied with ligation buffer containing ATP (10×)).
20. 10 mM dNTP mix.
21. Advantage® cDNA polymerase mix (50×) (Clontech; supplied with cDNA PCR reaction buffer (10×)).
22. Primer AdR-PCR: 5′-GGTCGCGGCCGAGGATC-3′ (10 μM).
23. QIAquick PCR Purification Kit (QIAGEN).
24. Agencourt AMPure XP Kit (Beckman Coulter) (optional, *see* below).
25. Agilent DNA 7500 kit (Agilent) (optional, *see* below).
26. TruSeq DNA HT Sample Prep Kit (Illumina) (optional, *see* below).

2.5 Solutions to Be Prepared

1. 3 M sodium acetate pH 5.2.
2. 5 M NaCl.
3. 70% (vol:vol) ethanol.
4. TENS buffer: 100 mM Tris–HCl pH 8.0, 5 mM EDTA pH 8.0, 200 mM NaCl, 0.2% SDS.

5. Lysis buffer: 5% sucrose (sterilized by filtering through a 0.22 μm filter), 100 mM Tris–HCl pH 9.1, 50 mM EDTA pH 8.0, 5% SDS.
6. Buffer A: 100 mM Tris–HCl pH 7.5, 100 mM EDTA pH 8.0, 100 mM NaCl, 0.5% SDS.
7. Buffer B: 1.43 M potassium acetate, 4.29 M LiCl.
8. 50 μM adapter AdR: Mix equal volumes (100–400 μL) of 100 μM oligonucleotides AdRt (5′-CTAATACGACTCACTA-TAGGGCAGCGTGGTCGCGGCCGAGGA-3′; should not be phosphorylated at the 5′ end) and AdRb (5′-TCCTCGGCCG-3′; should not be phosphorylated at the 5′ end) in a clean 1.5 mL tube. Place the tube in a beaker with a large volume (0.5–1 L) of boiling water and incubate for 5 min. Turn off heating and allow the beaker to cool to room temperature slowly (leave it for overnight). Make aliquots and store them at −20 °C.

2.6 Plasmids

1. p-attB-min.hsp70P-FRT-STOP#1-FRT-DamMyc[open] plasmid vector (Addgene plasmid #71809).
2. p-attB-min.hsp70P-FRT-STOP#1-FRT-DamMyc[closed] plasmid (Addgene plasmid #71810).

2.7 Fly Stocks

Flies can be raised on a standard cornmeal/molasses/agar food at 25 °C.

1. *y[1] M{vas-int.Dm}ZH-2A w[*]; M{3xP3-RFP.attP'}ZH-51C* (Bloomington stock #24482). This line contains the *vasa*-driven phiC31 integrase transgene on chromosome X and the phiC31 attP landing site associated with an RFP-coding transgene within cytogenetic region 51C.
2. *y[1] w[*]; M{w[+mC]=hs.min(FRT.STOP1)dam}ZH-51C* (Bloomington stock #65433). This line bears the STOP#1-Dam construct ("Dam only") under the control of the minimal *hsp70* promoter integrated into the 51C region and can be used to detect non-targeted Dam-dependent DNA methylation.
3. *y[1] w[*]; M{w[+mC]=hs-dam.4-HT-intein-L127C}ZH-51C* (Bloomington stock #65429) or *y[1] w[*]; M{w[+mC]=hs-dam.4-HT-intein-L127C-Lam}ZH-51C* (Bloomington stock #65430) or *y[1] w[*]; M{w[+mC]=hs-dam.4-HT-intein-L127C-Pc}ZH-51C* (Bloomington stock #65431). Control stocks to check for proper amplification of Dam methylated fragments.
4. Appropriate FLP driver line(s).

3 Methods

Unless otherwise specified, the procedures are performed at room temperature.

3.1 Testing Dam-POI Fusions in Cultured Cells

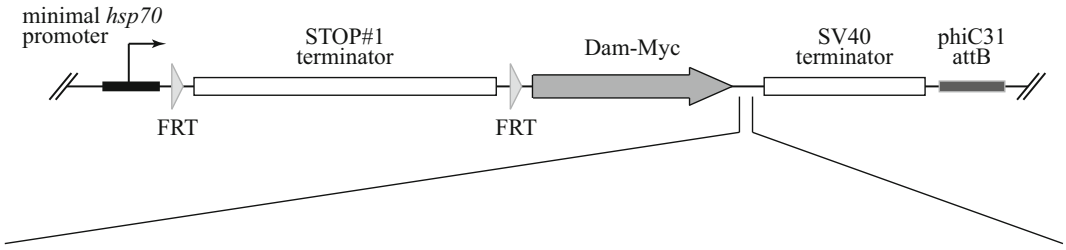
Because generation of transgenic flies is time-consuming, it is advisable to perform a simple preliminary test to ensure that either an N- or a C-terminal Dam fusion of the POI retains DNA adenine methyltransferase activity. The test can be done in cultured cells and requires only construction of plasmids encoding the Dam-POI proteins under the control of the full-length *hsp70* promoter (*see* **Note 1**). This experiment is optional, but can save a lot of time later.

1. To construct plasmids for constitutive expression of Dam-POI and POI-Dam fusion proteins, clone the POI-encoding DNA sequence into the pNDamMyc and pCMycDam vectors [13], respectively. Use the plasmid constructs to transfect cultured *Drosophila* Kc cells (*see* **Note 2**). Grow the cells for 24 h and then isolate genomic DNA and amplify the methylated GATC fragments. All these procedures have been described in detail earlier [18].
2. The successful amplification of Dam methylated GATC fragments does not always indicate that a meaningful (non-random) DamID profile is generated [50]. The nature of the amplified DNA fragments can be revealed either (a) by hybridization to an appropriate microarray (for the protocol, *see* [18]; even a stripped microarray would work for this purpose) or (b) by high-throughput sequencing (as described in Subheading 3.6; about 3–4 $\times$ 10^6 reads per replicate might be enough to understand whether the DamID profile makes biological sense). (c) Alternatively, if the association of the POI with a specific genomic location(s) is already known, methylation at individual GATC site(s) can be quantified by separate digestions with the DpnI and DpnII restriction enzymes, followed by qPCR (for the protocol, *see* [13, 22, 51]).

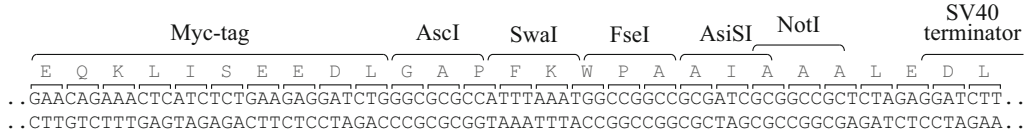
3.2 Generation of Transgenic Flies with FLP-Inducible Dam-POI Constructs

Depending on the available information, either an N- or a C-terminal Dam fusion of the POI, or even better both of them, can be used for DamID profiling in *Drosophila* tissues. The appropriate plasmid(s) should be constructed and integrated into a suitable genomic location (e.g., 51C) by phiC31 integrase-mediated recombination [52]. Transgenic fly generation might be carried out by a specialized company (e.g., BestGene, http://www.thebestgene.com).

1. To construct the plasmid for FLP-inducible expression of the Dam-POI fusion protein (N-terminal Dam fusion), use the p-attB-min.hsp70P-FRT-STOP#1-FRT-DamMyc[open] vector



p-attB-min.hsp70P-FRT-STOP#1-FRT-DamMyc[open] vector:



p-attB-min.hsp70P-FRT-STOP#1-FRT-DamMyc[closed] plasmid:

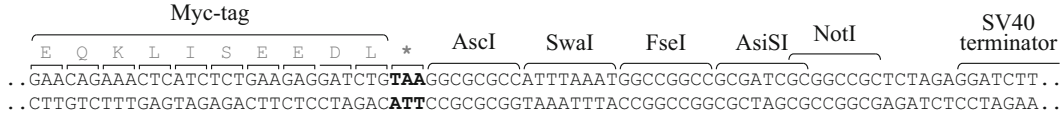


Fig. 3 The FLP-inducible Dam construct and possibilities for cloning an ORF of interest. The transgene consists of the ubiquitously active minimal *hsp70* promoter, the STOP#1 transcriptional terminator flanked by the FRT sites, the DNA sequence encoding Dam-Myc fusion protein, and the SV40 transcriptional terminator. In the p-attB-min.hsp70P-FRT-STOP#1-FRT-DamMyc[open] vector designed for generation of Dam-POI constructs, there are several unique restriction sites located downstream of the Dam-Myc ORF that are suitable for cloning. In the p-attB-min.hsp70P-FRT-STOP#1-FRT-DamMyc[closed] plasmid, which can be used for generating POI-Dam constructs, there is a stop codon immediately upstream of the polylinker

with a polylinker region containing a number of unique restriction sites [35] (Fig. 3). The construct for FLP-inducible expression of the POI-Dam protein (C-terminal Dam fusion) can be made by the isothermal assembly method [53]; the appropriate vector fragment can be amplified from the p-attB-min.hsp70P-FRT-STOP#1-FRT-DamMyc[closed] plasmid [35] (Fig. 3) (*see* **Note 3**). Verify the final structure of the plasmid(s) by DNA sequencing (*see* **Note 4**).

2. Use each plasmid construct to generate transgenic *Drosophila* lines by microinjection of the DNA in embryos of the *y[1] M{vas-int.Dm}ZH-2A w[*]; M{3xP3-RFP.attP'}ZH-51C* strain, in which a phiC31 attP landing site is located within cytogenetic region 51C (for the protocol, *see* [49]) (*see* **Notes 5** and **6**).
3. Verify the integrity of the transgene(s) by PCR with construct-specific primers from the genomic DNA isolated from each line. Although phiC31 integrase-mediated attP/attB recombination is an accurate process, in rare cases the transgene can be rearranged or truncated (*see* **Note 7**).

3.3 Preparation of Biological Samples for an FLP-Inducible DamID Experiment

To study the POI binding profile in a specific cell type, it is necessary to prepare and process a set of DNA samples isolated from the desired tissue (*see* Table 1). The experiment should consist of at least two biological replicates. Larvae or flies expressing the Dam-POI fusion protein or "Dam only" are obtained by crossing STOP#1-Dam-POI or control STOP#1-Dam flies with a line bearing an appropriate FLP driver.

1. Set up crosses between a strain carrying the chosen FLP driver and (a) flies carrying the STOP#1-Dam-POI transgene and (b) flies carrying the STOP#1-Dam transgene (Fig. 4a). If flies homozygous for the FLP or DamID transgene are not viable, use suitable balancer chromosomes to unambiguously identify the progeny that bears both transgenes.
2. Grow animals at 25 °C until they reach the desired developmental stage (*see* **Note 11**).

Table 1
Required experimental and control DNA samples

		No.	Genotype	Purpose
Experiment	Replicate 1	1	STOP#1-Dam-POI/ FLP driver	
		2	STOP#1-Dam/FLP driver	Correction of non-targeted Dam binding
	Replicate 2	3	STOP#1-Dam-POI/ FLP driver	
		4	STOP#1-Dam/FLP driver	Correction of non-targeted Dam binding
Controls		5	STOP#1-Dam-POI (*see* **Note 8**)	Control of the proper functionality of STOP#1 transcriptional terminator (there should be no leaky expression of the Dam-POI transgene)
		6	FLP driver (*see* **Note 8**)	Control for the absence of DNA degradation during extraction
		7	$Dam^{4\text{-HT-intein@L127C}}$ (*see* **Note 9**)	Positive control for the amplification of Dam methylated fragments
		8	STOP#1-Dam-POI / FLP driver (*see* **Note 10**)	"no DpnI": Control for the specificity of amplification of Dam methylated fragments
		9	$Dam^{4\text{-HT-intein@L127C}}$ (*see* **Note 9**)	"no DpnI": Control for the specificity of amplification of Dam methylated fragments
		10	STOP#1-Dam-POI/ FLP driver (*see* **Note 10**)	"no T4 DNA ligase": Control for the specificity of amplification of adapter-ligated DNA fragments
		11	$Dam^{4\text{-HT-intein@L127C}}$ (*see* **Note 9**)	"no T4 DNA ligase": Control for the specificity of amplification of adapter-ligated DNA fragments

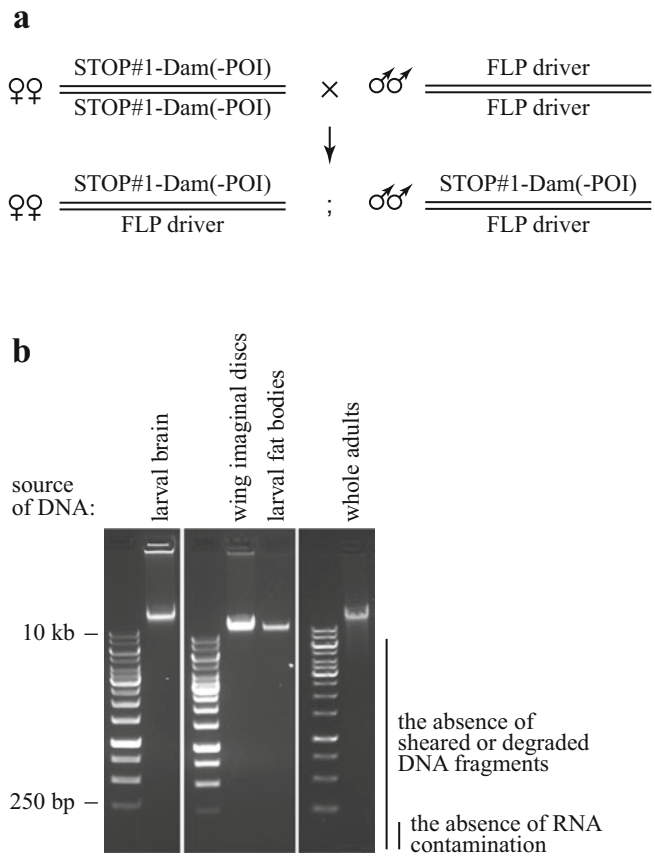


Fig. 4 Preparation of samples for amplification of Dam methylated genomic fragments. (**a**) A genetic cross between a line homozygous for a DamID construct [STOP#1-Dam-POI or STOP#1-Dam abbreviated as STOP#1-Dam(-POI)] and a line carrying an FLP driver activate DNA methylation in the cells of interest (in the example shown, both transgenes are assumed to be on the same chromosome, e.g., chromosome 2); the progeny from the depicted cross can be used to dissect the appropriate tissue. (**b**) Appearance of DNA isolated from different dissected larval tissues and from whole adults run on an agarose gel. The DNA appears as a single high-molecular weight band indicating the lack of degradation or shearing. Samples should also be free of RNA

3.4 Isolation of Genomic DNA

Described below are three protocols for isolation of genomic DNA from small *Drosophila* tissue samples (the first two protocols) or from whole flies (the third protocol). The first protocol is simple but works only for "soft" tissues such as brain dissected from wandering third instar larvae. The second protocol is optimized for DNA isolation from "hard" tissues (for example, larval fat bodies, wing imaginal discs, or salivary glands). This protocol might be also suitable for any other larval, pupal, or adult tissue, but the minimal amount of starting material should be determined in each case. The third protocol describes how to isolate genomic DNA from adult flies (or their parts, e.g., heads).

3.4.1 Isolation of DNA from "Soft" Tissues (Larval Brain)

1. For each biological replicate, 10–20 dissected larval brains (without associated imaginal discs) are required. Dissect the brains in PBS in a Pyrex spot plate under a binocular microscope using forceps and/or dissection needles. Collect the brains in 100 µL of TENS buffer in a clean 1.5 mL tube (*see* **Note 12**).
2. Centrifuge at 300 × *g* for 3 min to collect the material at the bottom of the tube, add 2 µL of proteinase K (from the 20 mg/mL stock), mix by tapping the tube, and incubate at 65 °C for 6 h. Mix the content of the tube each hour by tapping.
3. Add 1 µL of RNase A (from the 100 mg/mL stock) and incubate at 37 °C for 30 min.
4. Add 100 µL (1 volume) of phenol:chloroform:isoamyl alcohol (25:24:1 vol:vol:vol), mix well by shaking the tube, and centrifuge at 18,000 × *g* for 5 min.
5. Transfer 90 µL of the upper phase into a clean 1.5 mL tube and add 9 µL (0.1 volume) of 3 M sodium acetate pH 5.2, 270 µL (3 volumes) of 96% ethanol and 1 µL of glycogen (from the 20 mg/mL stock). Mix by inverting the tube several times and incubate at −20 °C overnight or at −80 °C for 30 min to precipitate DNA.
6. Centrifuge at 18,000 × *g* for 30 min at 4 °C, remove the supernatant, and wash the DNA pellet with 500 µL of ice-cold 70% (vol:vol) ethanol.
7. Centrifuge at 18,000 × *g* for 5 min at 4 °C, carefully remove the supernatant using a pipette, and air-dry the DNA pellet for 15 min.
8. Dissolve the DNA pellet in 12.5 µL of nuclease-free water, centrifuge at 18,000 × *g* for 5 min to collect undissolved material at the bottom of the tube, and transfer 12 µL of the solution into a clean 1.5 mL tube.
9. Analyze 2 µL of the isolated DNA sample on 1% agarose gel with 0.2 µg/mL of ethidium bromide to verify the integrity of DNA and the absence of RNA. To estimate DNA concentration, photograph the gel under UV light and compare fluorescence intensity of the DNA band with those of the DNA mass ladder run on the same gel (Fig. 4b); the concentration should be at least 100 ng/µL.
10. The DNA sample can be stored at −20 °C for up to several months.

3.4.2 Isolation of DNA from "Hard" Tissues (e.g., Larval Fat Bodies, Wing Imaginal Discs and Salivary Glands)

1. For each biological replicate, dissect in PBS in a Pyrex spot plate under a binocular microscope using forceps and/or dissection needles the following amount of tissue: 10 paired larval fat bodies, 30 paired larval salivary glands, or 40 wing imaginal discs. Collect the tissue in 100 μL of PBS in a clean 1.5 mL tube placed on ice (*see* **Note 12**).
2. Centrifuge at 300 × *g* for 3 min to collect the material at the bottom of the tube, add 400 μL of lysis buffer, 5 μL of proteinase K (from the 20 mg/mL stock), and disrupt the tissue by passing through 21G needle and a 1.0 mL syringe 20 times.
3. Incubate at 55 °C, 600 rpm for 4 h on a thermomixer.
4. Add 2 μL of RNase A (from the 100 mg/mL stock) and incubate at 37 °C for 30 min.
5. Add 500 μL (1 volume) of phenol:chloroform:isoamyl alcohol (25:24:1 vol:vol:vol), mix well by shaking the tube, and centrifuge at 18,000 × *g* for 5 min.
6. Transfer the upper phase into a clean 1.5 mL tube, and repeat **step 5** once and then proceed with **step 7**.
7. Transfer 380 μL of the upper phase into a clean 1.5 mL tube and add 38 μL (0.1 volume) of 5 M NaCl, 950 μL (2.5 volumes) of 96% ethanol, and 1 μL of glycogen (from the 20 mg/mL stock). Mix by inverting the tube several times and incubate at −20 °C overnight or at −80 °C for 30 min to precipitate DNA.
8. Centrifuge at 18,000 × *g* for 1 h at 4 °C, remove the supernatant and wash the DNA pellet with 500 μL of ice-cold 70% (vol:vol) ethanol (vortex until the pellet is loose).
9. Centrifuge at 18,000 × *g* for 15 min at 4 °C, carefully remove the supernatant using a pipette, and air-dry the DNA pellet for 15 min.
10. Dissolve the DNA pellet in 500 μL of nuclease-free water, transfer the solution into a Microcon-30 kDa centrifugal filter, and centrifuge at 14,000 × *g* for 12 min. This should reduce the volume to approximately 10 μL (sometimes, additional centrifugation for 1–2 min might be required).
11. Collect the DNA by placing the inverted Microcon-30 kDa centrifugal filter into a clean 1.5 mL tube and centrifuge at 1000 × *g* for 3 min.
12. Measure the volume of the solution and adjust it to 12 μL with nuclease-free water.
13. Perform **step 9** of the protocol 3.4.1; the DNA concentration should be at least 100 ng/μL.
14. The DNA sample can be stored at −20 °C for up to several months.

3.4.3 Isolation of DNA from Whole Adults

The first part of this protocol is based on the quick fly genomic DNA prep protocol of E.J. Rehm (http://www.fruitfly.org/about/methods/inverse.pcr.html). The protocol yields enough DNA to set up all necessary control reactions.

1. Collect 30–50 anesthetized adult flies in a clean 1.5 mL tube (*see* **Note 12**).
2. Add 200 μL of Buffer A and grind flies with a pellet pestle, add another 200 μL of Buffer A and continue grinding until only pieces of cuticle remain, and incubate at 65 °C for 1 h.
3. Add 800 μL of Buffer B, mix by inverting the tube, and keep on ice for 1 h.
4. Centrifuge at 18,000 × *g* for 15 min at 4 °C and transfer the supernatant (avoiding floating particles) into a clean 1.5 mL tube.
5. Centrifuge at 18,000 × *g* for 15 min at 4 °C to clean the sample from remaining pieces of the precipitate and transfer 1.0 mL of the supernatant into a clean 2.0 mL tube.
6. Add 700 μL (0.7 volume) of isopropanol and mix by inverting the tube several times.
7. Centrifuge at 18,000 × *g* for 15 min, carefully remove the supernatant using a pipette, and wash the DNA pellet with 500 μL of ice-cold 70% (vol:vol) ethanol.
8. Centrifuge at 18,000 × *g* for 5 min, carefully remove the supernatant using a pipette, and air-dry the DNA pellet for 15 min.
9. Dissolve the DNA pellet in 200 μL of nuclease-free water, centrifuge at 18,000 × *g* for 5 min to collect undissolved material at the bottom of the tube, and transfer 190 μL of the solution into a clean 1.5 mL tube.
10. Add 10 μL of SuRE/Cut Buffer H (10×) (*see* **Note 13**), 0.2 μL of RNase A (from the 100 mg/mL stock) and incubate at 37 °C for 30 min.
11. Add 200 μL (1 volume) of phenol:chloroform:isoamyl alcohol (25:24:1 vol:vol:vol), mix well by shaking the tube, and centrifuge at 18,000 × *g* for 5 min.
12. Transfer the upper phase into a clean 1.5 mL tube, add 200 μL (1 volume) of chloroform, mix well by shaking the tube, and centrifuge at 18,000 × *g* for 5 min.
13. Transfer 180 μL of the upper phase into a clean 1.5 mL tube and add 18 μL (0.1 volume) of 3 M sodium acetate pH 5.2 and 450 μL (2.5 volumes) of 96% ethanol. Mix by inverting the tube several times and incubate at −20 °C overnight or at −80 °C for 30 min to precipitate DNA.

14. Centrifuge at 18,000 × *g* for 30 min at 4 °C, remove the supernatant, and wash the DNA pellet with 500 μL of ice-cold 70% (vol:vol) ethanol.
15. Centrifuge at 18,000 × *g* for 10 min at 4 °C, carefully remove the supernatant using a pipette, and air-dry the DNA pellet for 15 min.
16. Dissolve the DNA pellet in 40 μL of nuclease-free water, centrifuge at 18,000 × *g* for 5 min to collect undissolved material at the bottom of the tube, and transfer 35 μL of the solution into a clean 1.5 mL tube.
17. Perform **step 9** of the protocol 3.4.1; the DNA concentration should be at least 200 ng/μL.
18. The DNA sample can be stored at −20 °C for up to several months.

3.5 Amplification of Dam Methylated GATC Fragments

To amplify the genomic fragments methylated at both ends at GATC sites by a Dam protein (Fig. 1), it is first necessary to digest DNA with the DpnI restriction enzyme, which cuts only Dam methylated but not unmethylated GATC sequences. Next, partially double-stranded DNA adapters are blunt-end ligated to DpnI digestion products (Fig. 5). Then, the digestion with the DpnII restriction enzyme destroys only fragments containing unmethylated GATC site(s) increasing the specificity of DamID mapping. After ligation of adapters to DpnI-digested DNA fragments, hemimethylated GATC sequences are generated, which are not subjected to DpnII digestion (Fig. 5). Finally, the GATC methylated fragments are amplified with an adapter-specific primer. All samples should be processed in parallel through all steps of the protocol. Use of filter tips is strongly recommended to avoid contamination of samples with other sources of Dam methylated DNA (e.g., plasmids).

1. For each sample (for the list of samples, *see* Subheading 3.3), combine the following components in a clean 0.2 mL PCR tube placed on ice: x μL of DNA (800 ng; *see* **Note 14**), 8.5–x μL of nuclease-free water, 1.0 μL of NEBuffer 4 (10×), and 0.5 μL of DpnI enzyme (from the 20 U/μL stock). For the "no DpnI" control reactions, add nuclease-free water instead of the enzyme. Mix by tapping the tubes.
2. Incubate in a PCR machine using the following program:
   (a) 37 °C: 6 h.
   (b) 80 °C: 20 min (to inactivate the enzyme).
   (c) 4–12 °C: up to overnight.
3. Place the tube on ice and add 6.2 μL of nuclease-free water, 0.8 μL of adapter AdR (from the 50 μM stock), 2.0 μL of

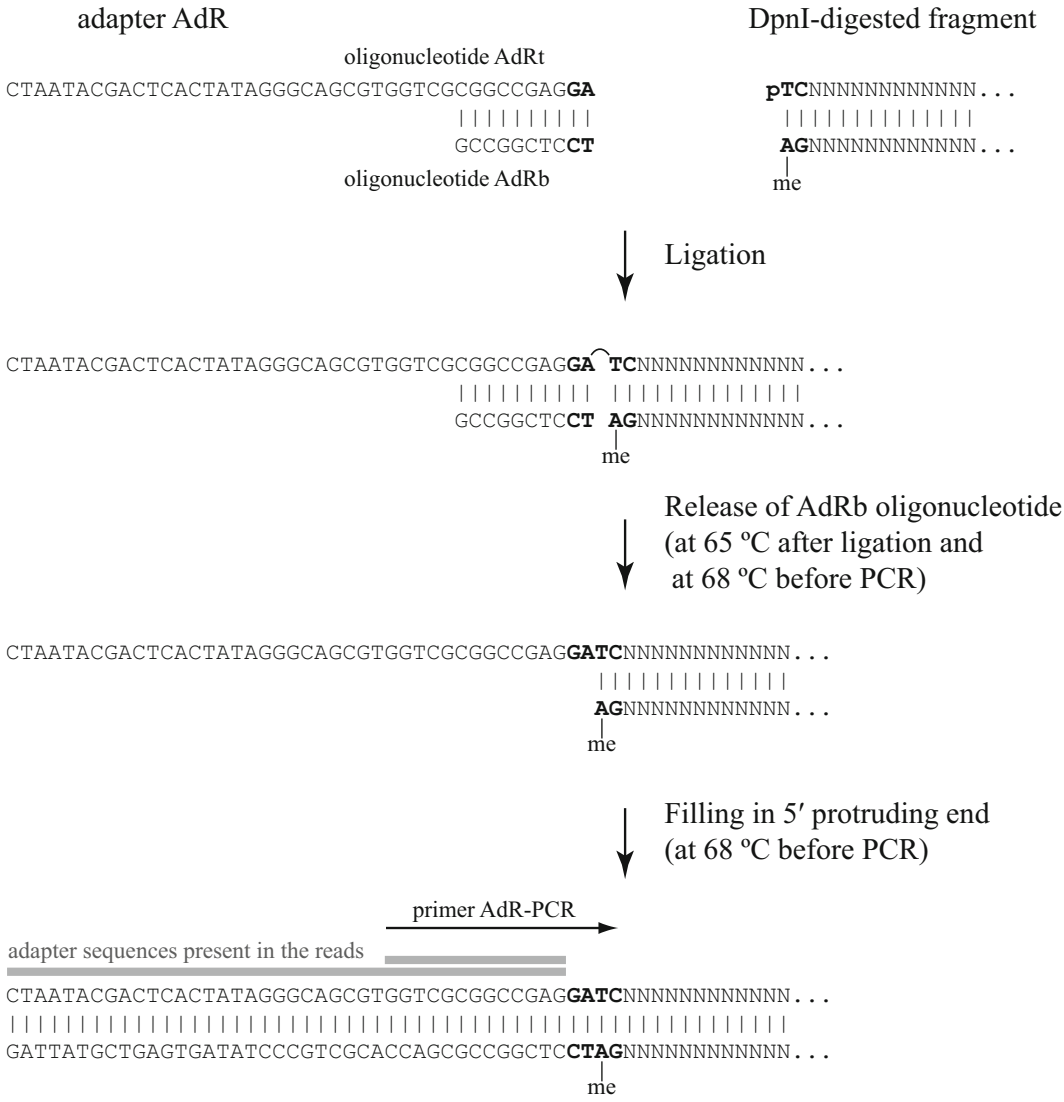


Fig. 5 The adapter AdR. The partially double-stranded adapter AdR contains GA nucleotides at its 3′ (blunt) end. Thus, the GATC sequence cleaved by the DpnI enzyme is restored after ligation of the adapter and 20–24 PCR steps. As the AdRb oligonucleotide is not phosphorylated at the 5′ end, only the top strand of the adapter becomes covalently linked to a DpnI-digested genomic fragment during the ligation reaction. Next, the bottom strand of the adapter is melt away and the 5′ protruding end of the DNA fragment is filled in by the polymerase activity. The short and long DamID adapter sequences that can be found in the reads, as well as the sequence of the primer AdR-PCR, are shown as horizontal *grey bars* and *black arrow*, respectively. me, methyl group

ligation buffer (10×), and 1.0 μL of T4 DNA ligase (from the 5 U/μL stock). For the "no T4 DNA ligase" control reactions, add nuclease-free water instead of the enzyme. Mix by tapping the tubes.

4. Incubate in a PCR machine using the following program:
   (a) 16 °C: 16 h.
   (b) 65 °C: 10 min (to inactivate the enzyme).
   (c) 4–12 °C: up to overnight.
5. Add 24.0 μL of nuclease-free water, 5.0 μL of NEBuffer DpnII (10×), and 1.0 μL of DpnII enzyme (from the 10 U/μL stock). Mix by tapping the tubes.
6. Incubate in a PCR machine using the following program:
   (a) 37 °C: 1 h.
   (b) 4–12 °C: up to overnight.
7. Transfer 10.0 μL of each sample (*see* **Note 15**) into a clean 0.2 mL PCR tube placed on ice and add the following components: 26.8 μL of nuclease-free water, 1.0 μL of 10 mM dNTP mix, 6.2 μL of primer AdR-PCR (from the 10 μM stock), 5.0 μL of cDNA PCR reaction buffer (10×), and 1.0 μL of Advantage® cDNA polymerase mix (50×). Mix by tapping the tubes.
8. Incubate in a PCR machine using the following program:
   (a) 68 °C: 10 min (to fill in 5′ protruding ends of the ligated adapters; Fig. 5).
   (b) 94 °C: 1 min.
   (c) 65 °C: 5 min.
   (d) 68 °C: 15 min.
   (e) 94 °C: 1 min.
   (f) 65 °C: 1 min.
   (g) 68 °C: 10 min.
   (h) Go to (e) 4×.
   (i) 94 °C: 1 min.
   (j) 65 °C: 1 min.
   (k) 68 °C: 2 min.
   (l) Go to (i) 13–15–17× (*see* **Note 16**).
   (m) 4–12 °C: up to overnight.
9. Analyze 10 μL (i.e., 1/5th) of each reaction on 1% agarose gel with 0.2 μg/mL of ethidium bromide. A successful experiment should result in a smear of PCR products in the range of ~200–2000 bp in samples prepared from tissues of larvae/flies bearing both the FLP driver and DamID construct, but not from all other (control) samples (Fig. 6).
10. The PCR products can be stored at −20 °C for at least 1 year.

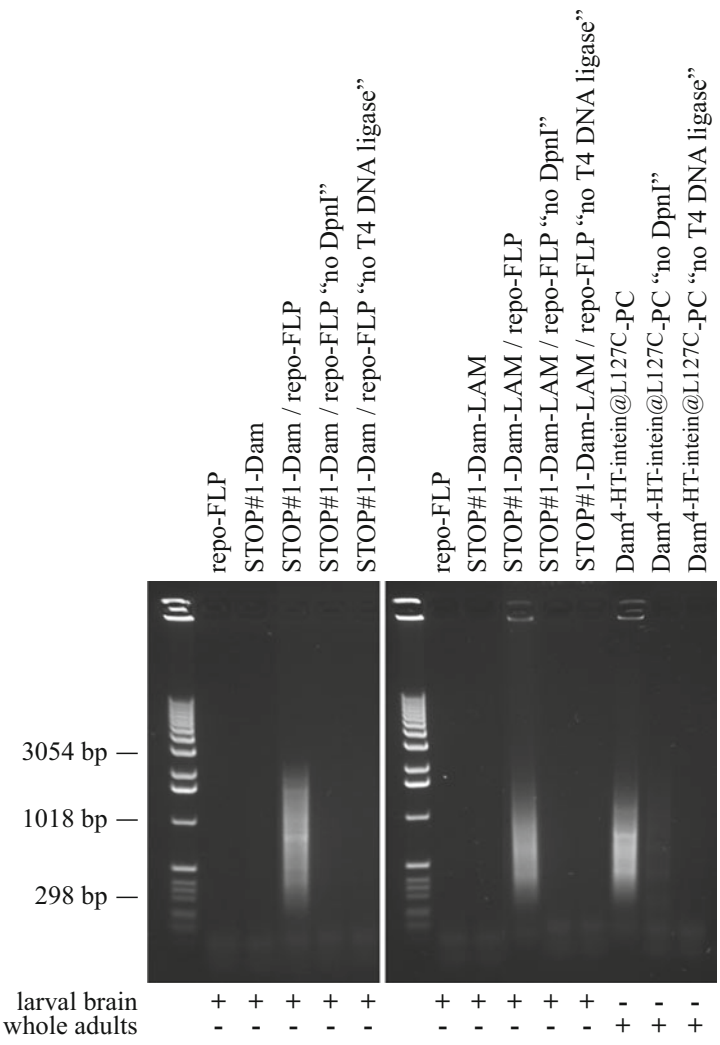


Fig. 6 A typical example of PCR-amplified Dam methylated DNA fragments analyzed on an agarose gel. A repo-FLP driver [39] was used to activate DamID transgenes in glial cells and DNA was isolated from whole larval brains. A smear of PCR products in the range of 0.2–2 kb is detected in experimental but not in control samples. The bands sometimes observed within the smear of PCR products could originate either from mitochondrial DNA [58] or repetitive elements of the genome

3.6 Library Preparation and High-Throughput Sequencing of Amplified DNA Fragments

Only experimental (but not control) samples should be subjected to high-throughput sequencing. During the preparation of samples for sequencing, the amplified DNA fragments are randomly sheared to ensure that long PCR products will not be underrepresented or lost during the Illumina cluster generation and sequencing steps. Typically, the size of the DNA fragments is reduced to a range of 100–500 bp with a peak around 300 bp. The **steps 2–5** of the following protocol are usually performed by a DNA sequencing facility.

1. For each sample, purify the amplified DNA fragments using QIAquick PCR Purification Kit according to the manufacturer's instructions. Elute the DNA in 100 μL of nuclease-free water. Use 1 μL to measure the DNA concentration with a NanoDrop spectrophotometer; the typical concentration should be ~50–100 ng/μL of DNA (*see* **Note 17**).
2. Dilute 2–3 μg of the purified PCR products into 100 μL of nuclease-free water, transfer into a clean snap-cap microTUBE (6 × 16 mm) with AFA fiber, and shear using an S2 ultrasonicator with a duty cycle of 10%, intensity of 5, cycles/burst of 200 for a total of 45 s, bath at 4 °C.
3. Purify the DNA using 160 μL of Agencourt AMPure magnetic beads according to the manufacturer's instructions. Elute the DNA in 50 μL of nuclease-free water and use 1 μL to monitor the size distribution of DNA fragments and measure the DNA concentration on an Agilent Bioanalyzer 2100 using a DNA 7500 Kit chip.
4. Use 1 μg of DNA to prepare an indexed Illumina sequencing library with TruSeq DNA HT Sample Prep Kit according to the manufacturer's instructions.
5. Combine equal amount of differently indexed libraries together and sequence them on a HiSeq instrument with 50–100 bp single-end reads according to the manufacturer's instructions. Usually 30–40 × 10^6 reads per sample (Dam-POI or Dam replicate) are sufficient to produce the DamID profile of high quality.

3.7 High-Throughput Sequencing Data Processing

The processing of DamID-seq data has been recently described in detail [54–56]. Thus, here, I only highlight some features that are relevant for the FLP-inducible DamID-seq data analysis.

1. It should be noted that two variants of the DamID adapter sequence can be found in the reads: (a) the most frequent short adapter (5′-GGTCGCGGCCGAG-3′) and (b) the less frequent long adapter (5′-CTAATACGACTCACTATAGGG-CAGCGTGGTCGCGGCCGAG-3′); they correspond to the truncated sequences (without the GA nucleotides at the 3′ ends) of the primer AdR-PCR and oligonucleotide AdRt used for the amplification and ligation steps, respectively (Fig. 5). This is due to the absence of DNA purification step between the ligation of adapters and PCR.
2. Typically, 30–60% of reads obtained from the Dam-POI or "Dam only" sample can be uniquely mapped to the *Drosophila* genome and thus used for the subsequent analysis. However, in the case of DamID mapping of heterochromatin proteins (which are mostly associated with repetitive DNA sequences), the fraction of such reads can be very low, less than 2%.

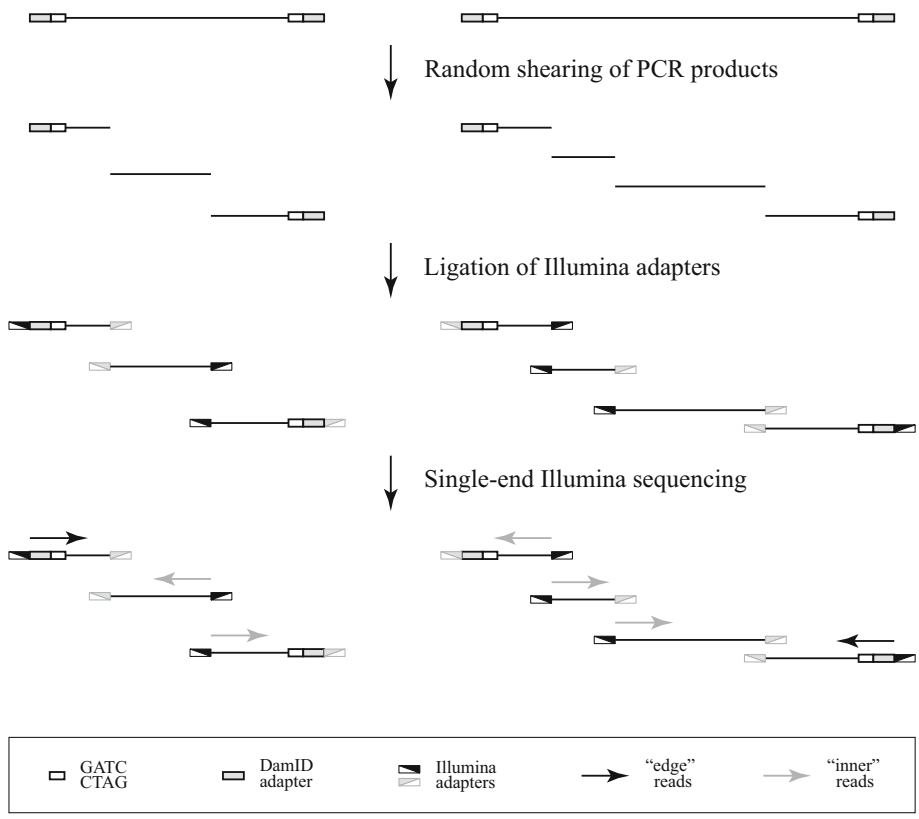


Fig. 7 Preparation of DamID products for high-throughput sequencing. Purified DNA fragments are randomly sheared, end repaired, A-tailed and ligated to Illumina adapters. After a few rounds of indexing PCR, the resulting libraries of DNA fragments obtained from different DamID samples are pooled and subjected to high-throughput single-end sequencing. "Edge," but not "inner" reads start with the sequence of the DamID adapter (which sometimes is truncated at the 5′ end, due to the DNA shearing step) followed by the GATC motif

3. The shearing of the PCR products during the library preparation step has an important consequence for data analysis. Namely, only part of the reads starts with the GATC sequence (after trimming of the adapters) (Fig. 7). Such ("edge") reads are undoubtedly from Dam methylated GATC fragments. Other ("inner") reads, without a GATC sequence at their beginning, could come from either (a) internal parts of Dam methylated GATC fragments or (b) random genomic fragments generated by DNA shearing during extraction from dissected tissue. The ratio between "edge" and "inner" reads is usually between 1:2.5 and 1:3. If desired, only "edge" reads can be used to generate the DamID profile. However, since typically the noise (amount of PCR products in the control samples) is very low, it is worthwhile to process "edge" and "inner" reads together to build a profile based on a large read counts.

4 Notes

1. In principle, constructs with the minimal *hsp70* promoter (*see* Subheading 3.2) can be used, but this has not been tested.
2. To avoid artifacts, a POI-expressing cell line should be used.
3. FLP-inducible vectors with the full-length (instead of the minimal) *hsp70* promoter are available and can be obtained from the author. Such vectors allow an increase of the expression level of Dam proteins by heat shock treatments, which might be required for particular cell types or tissues.
4. It is important to sequence exactly the same plasmid preparation that will be used later for *Drosophila* transgenesis; this will avoid a chance of random mutagenesis during plasmid maintenance and propagation in *E. coli*.
5. The 51C insertion locus is characterized by a relatively low expression level of the integrated transgene [52] and has been previously used for DamID experiments [35, 57]. Importantly, the *Drosophila* line with the FLP-inducible "Dam only" control construct inserted into the 51C locus already exists [35]. It is advisable to establish a few independent lines carrying the Dam-POI expressing construct, as the chromosome that contains this transgene might acquire second site mutations unrelated with the insertion site. Whenever possible, use homozygous viable lines for the experiments.
6. If another attP landing site is chosen for integration of the Dam-POI construct(s), then the control "Dam only" construct (the p-attB-min.hsp70P-FRT-STOP#1-FRT-DamMyc [closed] plasmid) should be integrated into the same genomic location to avoid position effects on gene expression. Since the functionality of the transcriptional terminator present in the FLP-inducible DamID constructs might be affected by the local chromatin structure at the chosen attP landing site, special attention should be paid to the appropriate control during the DamID experiment (*see* Subheading 3.3).
7. Since the DamID transgene is expressed (after its FLP-mediated activation) at a low level, assessing the functionality of the Dam-POI by a rescue experiment of a mutation in the POI-encoding gene may be problematic. In addition, a ubiquitous activation of the transgene (e.g., by FLP expressed under the control of the *hsp70* promoter) could lead to lethality at some developmental stage [35].
8. Usually one replicate of this control reaction is enough.
9. $Dam^{4\text{-HT-intein@L127C}}$ is a hypomorphic Dam mutant [35]. The genomic DNA of flies carrying the $Dam^{4\text{-HT-intein@L127C}}$ transgene is methylated to some extent without activation of the

transgene. For this control, DNA should be isolated from adult flies (*see* Subheading 3.4.3). In our experience, $Dam^{4\text{-HT-intein@L127C}}$-Lamin Dm0 ($Dam^{4\text{-HT-intein@L127C}}$-LAM) or $Dam^{4\text{-HT-intein@L127C}}$-Polycomb ($Dam^{4\text{-HT-intein@L127C}}$-PC) flies work even better, because they exhibit a slightly higher level of DNA methylation than $Dam^{4\text{-HT-intein@L127C}}$ alone.

10. Usually there is no need to perform this control reaction for each individual experimental sample. Choose any of the replicates or even use a mix of them.
11. Experiment can be performed at any other desired temperature. The minimal *hsp70* promoter present in the FLP-inducible DamID transgenes is not sensitive to heat shock.
12. The sample can be stored at −20 °C for up to several months and thawed on ice just prior to use.
13. RNase A is known to be active under a wide range of reaction conditions, allowing the use of another buffer for restriction digestion.
14. Robust results can be obtained with 0.5–1 μg of DNA.
15. The remaining 40.0 μL of each sample can be stored at −20 °C for up to 1 year.
16. To find out the optimal number of PCR cycles required, divide each reaction in three equal parts and run them using three different amplification programs: 13×, 15× and 17×. PCR products should be observed in experimental samples, but not in the controls.
17. The purified DNA can be stored at −20 °C for at least 1 year.

Acknowledgments

I thank the laboratory of Prof. B. van Steensel at the Netherlands Cancer Institute (Amsterdam, the Netherlands) for providing an excellent working environment, the NKI Genomics Core Facility for help with the development of the protocol for preparation of DNA samples for Illumina high-throughput sequencing; Anna A. Ogienko for technical support; Mario Amendola, Maurizio Gatti, and the members of Laboratory of Cell Division for critical reading of the manuscript and helpful suggestions. This work was supported by the grant of Russian Science Foundation no. 16-14-10288.

References

1. Bartman CR, Blobel GA (2015) Perturbing chromatin structure to understand mechanisms of gene expression. Cold Spring Harb Symp Quant Biol 80:207–212
2. Allis CD, Jenuwein T (2016) The molecular hallmarks of epigenetic control. Nat Rev Genet 17(8):487–500

3. Khurana S, Oberdoerffer P (2015) Replication stress: a lifetime of epigenetic change. Genes (Basel) 6(3):858–877
4. Dabin J, Fortuny A, Polo SE (2016) Epigenome maintenance in response to DNA damage. Mol Cell 62(5):712–727
5. van Steensel B (2005) Mapping of genetic and epigenetic regulatory networks using microarrays. Nat Genet 37 Suppl:S18–S24
6. Southall TD, Brand AH (2007) Chromatin profiling in model organisms. Brief Funct Genomic Proteomic 6(2):133–140
7. Aughey GN, Southall TD (2016) Dam it's good! DamID profiling of protein–DNA interactions. Wiley Interdiscip Rev Dev Biol 5:25–37
8. Furey TS (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. Nat Rev Genet 13(12):840–852
9. Christova R (2013) Detecting DNA–protein interactions in living cells–ChIP approach. Adv Protein Chem Struct Biol 91:101–133
10. Rodriguez-Ubreva J, Ballestar E (2014) Chromatin immunoprecipitation. Methods Mol Biol 1094:309–318
11. Rhee HS, Pugh BF (2011) Comprehensive genome-wide protein–DNA interactions detected at single-nucleotide resolution. Cell 147(6):1408–1419
12. Wardle FC, Tan H (2015) A ChIP on the shoulder? Chromatin immunoprecipitation and validation strategies for ChIP antibodies. F1000Res 4:235
13. van Steensel B, Henikoff S (2000) Identification of *in vivo* DNA targets of chromatin proteins using tethered dam methyltransferase. Nat Biotechnol 18(4):424–428
14. van Steensel B, Delrow J, Henikoff S (2001) Chromatin profiling using targeted DNA adenine methyltransferase. Nat Genet 27 (3):304–308
15. Ratel D, Ravanat JL, Berger F, Wion D (2006) N6-methyladenine: the other methylated base of DNA. BioEssays 28(3):309–315
16. Shagin DA, Lukyanov KA, Vagner LL, Matz MV (1999) Regulation of average length of complex PCR product. Nucleic Acids Res 27 (18):e23
17. Choksi SP, Southall TD, Bossing T, Edoff K, de Wit E, Fischer BE, van Steensel B, Micklem G, Brand AH (2006) Prospero acts as a binary switch between self-renewal and differentiation in *Drosophila* neural stem cells. Dev Cell 11 (6):775–789
18. Greil F, Moorman C, van Steensel B (2006) DamID: mapping of *in vivo* protein–genome interactions using tethered DNA adenine methyltransferase. Methods Enzymol 410:342–359
19. Coffin SR, Reich NO (2009) *Escherichia coli* DNA adenine methyltransferase: intrasite processivity and substrate-induced dimerization and activation. Biochemistry 48 (31):7399–7410
20. Pollak AJ, Reich NO (2012) Proximal recognition sites facilitate intrasite hopping by DNA adenine methyltransferase: mechanistic exploration of epigenetic gene regulation. J Biol Chem 287(27):22873–22881
21. Lu L, Patel H, Bissler JJ (2002) Optimizing DpnI digestion conditions to detect replicated DNA. BioTechniques 33(2):316–318
22. Germann S, Juul-Jensen T, Letarnec B, Gaudin V (2006) DamID, a new tool for studying plant chromatin profiling *in vivo*, and its use to identify putative LHP1 target loci. Plant J 48(1):153–163
23. Venkatasubrahmanyam S, Hwang WW, Meneghini MD, Tong AH, Madhani HD (2007) Genome-wide, as opposed to local, antisilencing is mediated redundantly by the euchromatic factors Set1 and H2A.Z. Proc Natl Acad Sci U S A 104(42):16609–16614
24. Woolcock KJ, Gaidatzis D, Punga T, Buhler M (2011) Dicer associates with chromatin to repress genome activity in *Schizosaccharomyces pombe*. Nat Struct Mol Biol 18(1):94–99
25. Gonzalez-Aguilera C, Ikegami K, Ayuso C, de Luis A, Iniguez M, Cabello J, Lieb JD, Askjaer P (2014) Genome-wide analysis links emerin to neuromuscular junction activity in *Caenorhabditis elegans*. Genome Biol 15(2):R21
26. Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, van Steensel B (2010) Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. Cell 143(2):212–224
27. Peric-Hupkes D, Meuleman W, Pagie L, Bruggeman SW, Solovei I, Brugman W, Graf S, Flicek P, Kerkhoven RM, van Lohuizen M, Reinders M, Wessels L, van Steensel B (2010) Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol Cell 38(4):603–613
28. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, van Steensel B (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453(7197):948–951

29. Vogel MJ, Peric-Hupkes D, van Steensel B (2007) Detection of *in vivo* protein–DNA interactions using DamID in mammalian cells. Nat Protoc 2(6):1467–1478
30. Kind J, Pagie L, de Vries SS, Nahidiazar L, Dey SS, Bienko M, Zhan Y, Lajoie B, de Graaf CA, Amendola M, Fudenberg G, Imakaev M, Mirny LA, Jalink K, Dekker J, van Oudenaarden A, van Steensel B (2015) Genome-wide maps of nuclear lamina interactions in single human cells. Cell 163(1):134–147
31. Deal RB, Henikoff S (2010) A simple method for gene expression and chromatin profiling of individual cell types within a tissue. Dev Cell 18 (6):1030–1040
32. Bonn S, Zinzen RP, Girardot C, Gustafson EH, Perez-Gonzalez A, Delhomme N, Ghavi-Helm Y, Wilczynski B, Riddell A, Furlong EE (2012) Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat Genet 44(2):148–156
33. Southall TD, Gold KS, Egger B, Davidson CM, Caygill EE, Marshall OJ, Brand AH (2013) Cell-type-specific profiling of gene expression and chromatin binding without cell isolation: assaying RNA Pol II occupancy in neural stem cells. Dev Cell 26(1):101–112
34. Marshall OJ, Southall TD, Cheetham SW, Brand AH (2016) Cell-type-specific profiling of protein–DNA interactions without cell isolation using targeted DamID with next-generation sequencing. Nat Protoc 11 (9):1586–1598
35. Pindyurin AV, Pagie L, Kozhevnikova EN, van Arensbergen J, van Steensel B (2016) Inducible DamID systems for genomic mapping of chromatin proteins in *Drosophila*. Nucleic Acids Res 44(12):5646–5657
36. Struhl G, Basler K (1993) Organizing activity of wingless protein in *Drosophila*. Cell 72 (4):527–540
37. Qi J, Su S, McGuffin ME, Mattox W (2006) Concentration dependent selection of targets by an SR splicing regulator results in tissue-specific RNA processing. Nucleic Acids Res 34(21):6256–6263
38. Lakso M, Sauer B, Mosinger B Jr, Lee EJ, Manning RW, Yu SH, Mulder KL, Westphal H (1992) Targeted oncogene activation by site-specific recombination in transgenic mice. Proc Natl Acad Sci U S A 89(14):6232–6236
39. Silies M, Yuva Y, Engelen D, Aho A, Stork T, Klambt C (2007) Glial cell migration in the eye disc. J Neurosci 27(48):13130–13139
40. Millard SS, Flanagan JJ, Pappu KS, Wu W, Zipursky SL (2007) Dscam2 mediates axonal tiling in the *Drosophila* visual system. Nature 447(7145):720–724
41. Therrien M, Wong AM, Rubin GM (1998) CNK, a RAF-binding multidomain protein required for RAS signaling. Cell 95 (3):343–353
42. Narbonne K, Besse F, Brissard-Zahraoui J, Pret AM, Busson D (2004) *polyhomeotic* is required for somatic cell proliferation and differentiation during ovarian follicle formation in *Drosophila*. Development 131(6):1389–1400
43. Lee CH, Herman T, Clandinin TR, Lee R, Zipursky SL (2001) N-cadherin regulates target specificity in the *Drosophila* visual system. Neuron 30(2):437–450
44. Nern A, Pfeiffer BD, Rubin GM (2015) Optimized tools for multicolor stochastic labeling reveal diverse stereotyped cell arrangements in the fly visual system. Proc Natl Acad Sci U S A 112(22):E2967–E2976
45. Bohm RA, Welch WP, Goodnight LK, Cox LW, Henry LG, Gunter TC, Bao H, Zhang B (2010) A genetic mosaic approach for neural circuit mapping in *Drosophila*. Proc Natl Acad Sci U S A 107(37):16378–16383
46. Pickersgill H, Kalverda B, de Wit E, Talhout W, Fornerod M, van Steensel B (2006) Characterization of the *Drosophila melanogaster* genome at the nuclear lamina. Nat Genet 38 (9):1005–1014
47. Luo SD, Shi GW, Baker BS (2011) Direct targets of the *D. melanogaster* DSX^F protein and the evolution of sexual development. Development 138(13):2761–2771
48. Stocker H, Gallant P (2008) Getting started : an overview on raising and handling *Drosophila*. Methods Mol Biol 420:27–44
49. Ringrose L (2009) Transgenesis in *Drosophila melanogaster*. Methods Mol Biol 561:3–19
50. van Bemmel JG, Filion GJ, Rosado A, Talhout W, de Haas M, van Welsem T, van Leeuwen F, van Steensel B (2013) A network model of the molecular organization of chromatin in *Drosophila*. Mol Cell 49(4):759–771
51. Kind J, Pagie L, Ortabozkoyun H, Boyle S, de Vries SS, Janssen H, Amendola M, Nolen LD, Bickmore WA, van Steensel B (2013) Single-cell dynamics of genome-nuclear lamina interactions. Cell 153(1):178–192
52. Bischof J, Maeda RK, Hediger M, Karch F, Basler K (2007) An optimized transgenesis system for *Drosophila* using germ-line-specific phiC31 integrases. Proc Natl Acad Sci U S A 104(9):3312–3317
53. Gibson DG (2011) Enzymatic assembly of overlapping DNA fragments. Methods Enzymol 498:349–361

54. Marshall OJ, Brand AH (2015) damidseq_pipeline: an automated pipeline for processing DamID sequencing datasets. Bioinformatics 31(20):3371–3373
55. Li R, Hempel LU, Jiang T (2015) A non-parametric peak calling algorithm for DamID-Seq. PLoS One 10(3):e0117415
56. Gomez-Saldivar G, Meister P, Askjaer P (2016) DamID analysis of nuclear organization in *Caenorhabditis elegans*. Methods Mol Biol 1411:341–358
57. Tolhuis B, Blom M, Kerkhoven RM, Pagie L, Teunissen H, Nieuwland M, Simonis M, de Laat W, van Lohuizen M, van Steensel B (2011) Interactions among Polycomb domains are guided by chromosome architecture. PLoS Genet 7(3):e1001343
58. Robson MI, Schirmer EC (2016) The application of DamID to identify peripheral gene sequences in differentiated and primary cells. Methods Mol Biol 1411:359–386

Chapter 8

DNA Methylation Profiling Using Long-Read Single Molecule Real-Time Bisulfite Sequencing (SMRT-BS)

Yao Yang and Stuart A. Scott

Abstract

For the past two decades, bisulfite sequencing has been a widely used method for quantitative CpG methylation detection of genomic DNA. Coupled with PCR amplicon cloning, bisulfite Sanger sequencing allows for allele-specific CpG methylation assessment; however, its time-consuming protocol and inability to multiplex has recently been overcome by next-generation bisulfite sequencing techniques. Although high-throughput sequencing platforms have enabled greater accuracy in CpG methylation quantitation as a result of increased bisulfite sequencing depth, most common sequencing platforms generate reads that are similar in length to the typical bisulfite PCR size range (~300–500 bp). Using the Pacific Biosciences (PacBio) sequencing platform, we developed single molecule real-time bisulfite sequencing (SMRT-BS), which is an accurate targeted CpG methylation analysis method capable of a high degree of multiplexing and long read lengths. SMRT-BS is reproducible and was found to be concordant with other lower throughput quantitative CpG methylation methods. Moreover, the ability to sequence up to ~1.5–2.0 kb amplicons, when coupled with an optimized bisulfite-conversion protocol, allows for more thorough assessment of CpG islands and increases the capacity for studying the relationship between single nucleotide variants and allele-specific CpG methylation.

Key words DNA methylation, CpG islands, SMRT sequencing, Bisulfite sequencing, PacBio sequencing, Long-read sequencing, Multiplex DNA methylation analysis

1 Introduction

Since the initial report on bisulfite Sanger sequencing in 1992 [1], the technique has been widely used for DNA methylation discovery and as a diagnostic assay for detecting CpG methylation abnormalities at imprinting control regions and across specific CpG islands [2, 3]. However, a major limitation of bisulfite Sanger sequencing is the need for PCR amplicon cloning, which translates to a laborious protocol and an inability to multiplex distinct amplicons. Bisulfite pyrosequencing is a faster, reproducible, and quantitative analysis of DNA methylation, but is restricted to short read lengths (~150 bp) and also has limited capacity for multiplexing [4]. In contrast to these targeted

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_8,

bisulfite sequencing techniques, reduced representation bisulfite sequencing [5] and whole-genome bisulfite sequencing [6] can quantitatively profile CpG methylation across an entire genome in a single experiment. However, these genome-wide approaches require significant computational expertise and infrastructure, and may be prohibitively expensive given their low throughput.

To address the need for a quantitative and highly multiplexed targeted bisulfite sequencing method capable of long read lengths, we recently developed a technique that combines bisulfite conversion with third-generation single molecule real-time (SMRT) sequencing using the Pacific Biosciences (PacBio) platform [7]. Coupled with an optimized long-range bisulfite amplification protocol and empowered by the long read lengths of SMRT sequencing (averaging ~10–15 kb) [8], SMRT bisulfite sequencing (SMRT-BS) can accurately measure CpG methylation across multiplexed ~1.5 kb regions without the need for PCR amplicon subcloning [7].

2 Materials

2.1 Bisulfite Conversion

Methylamp™ DNA Modification Kit (Epigentek) (*see* **Note 1**).

2.2 Bisulfite PCR and Amplicon Purification

Premix Taq DNA Polymerase Hot Start Version (Takara).
ASI Agarose (Alkali Scientific).
QiaQuick PCR Purification Kit (Qiagen).
RNA6000 pico kit (Agilent).
1 kb DNA Ladder (Thermo Fisher).
NanoDrop 1000 (Thermo Scientific).

2.3 SMRT Sequencing

SMRTbell™ Template Prep Kit (Pacific Biosciences).
DNA/Polymerase Binding Kit (Pacific Biosciences).
MagBead Kit for amplicons >1 kb (Pacific Biosciences).
DNA Sequencing Reagent (Pacific Biosciences).
DNA Internal Control Complex (Pacific Biosciences).
SMRT® Cells (Pacific Biosciences).
AMPure® PB beads (Pacific Biosciences).

3 Methods

The overall workflow of SMRT-BS is illustrated in Fig. 1, which can be separated into five general steps: (1) bisulfite conversion of genomic DNA; (2) first round PCR amplification of bisulfite-treated DNA using region-specific primers coupled with universal oligonucleotide tags; (3) re-amplification of amplicon templates using universal anti-tag primers coupled with sample-specific barcodes; (4) amplicon pooling and SMRT sequencing; and (5) data analysis and CpG methylation quantitation.

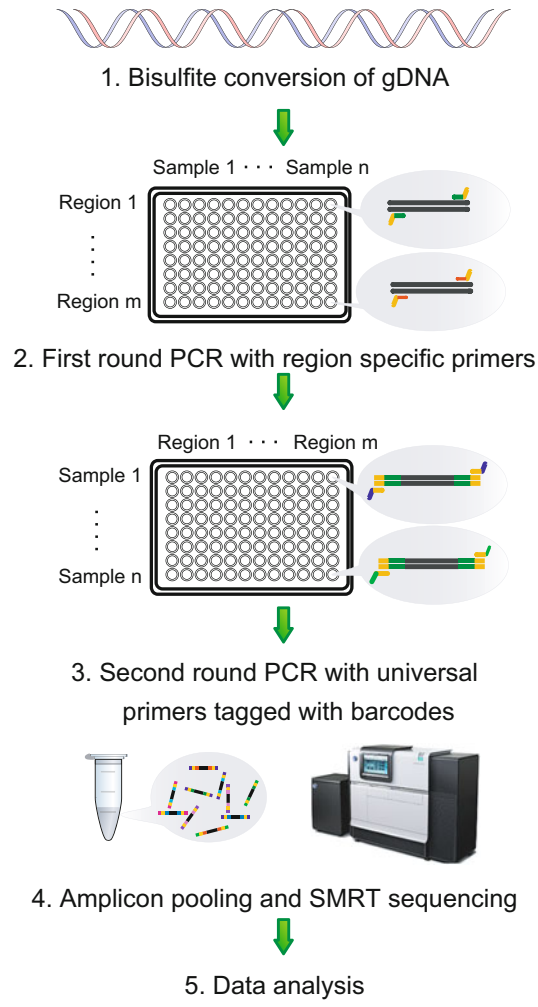


Fig. 1 Illustration of the SMRT bisulfite sequencing (SMRT-BS) workflow. Image adapted from Yang et al., *BMC Genomics*, 2015 [7]

3.1 Bisulfite Conversion

Although a number of bisulfite conversion kits are commercially available, our initial validation of SMRT-BS employed the Methylamp™ DNA Modification Kit (Epigentek) [7], which was performed according to the manufacturer's instructions.

1. For each sample, place a 1.7 ml tube on a rack and add 500 ng–1 μg genomic DNA to each tube (*see* **Note 2**).
2. Follow the Methylamp™ DNA Modification Kit protocol and elute the bisulfite-converted DNA with 20 μl of elution buffer.
3. (Optional) Modified DNA can be quantified by a Nanodrop 1000 (Thermo Scientific) using the ssDNA application and sized with the Agilent Bioanalyzer 2100 using the RNA6000 pico kit.

3.2 Primer Design

1. For the first round PCR amplification of bisulfite converted DNA, design region-specific forward and reverse primers using MethPrimer (http://www.urogene.org/methprimer) or a comparable program capable of designing bisulfite sequencing PCR primers. Add universal oligonucleotide tags to the 5′ end of both forward and reverse primers prior to synthesis to enable the addition of barcodes through a second round PCR.
2. For the second round PCR amplification, add sample-specific barcodes to the 5′ end of the universal oligonucleotide anti-tag primers (*see* **Note 3**).

3.3 Bisulfite PCR Amplification

First round PCR amplifies bisulfite-converted DNA using region-specific forward and reverse primers that have 5′ universal oligonucleotide tags for subsequent barcoding.

1. Place a 96 well PCR plate on ice for PCR preparation.
2. Prepare master mix for the reaction in a 1.7 ml tube. For each 20 μl reaction, add 11.8 μl H_2O, 2 μl 10× PCR buffer with Mg^{2+}, 3.2 μl dNTP (1.25 mM each) and 0.2 μl TaKaRa Taq HS.
3. Dispense 17.2 μl of master mix to each well.
4. Add 0.8 μl of primer mix (5 μM each) and 2 μl of bisulfite-converted DNA to each well.
5. Seal the plate with an adhesive film, vortex, and centrifuge to bring all components to the bottom of the plate.
6. Place the PCR plate in a thermal cycler and set the amplification program as following: initial denaturation step at 94 °C for 2 min followed by 35 amplification cycles (94 °C for 20 s, annealing temperature for 45 s, and 65 °C for 1 min/kb + 30 s) and a final incubation at 65 °C for 5 min (*see* **Note 4**).
7. Electrophorese PCR products through a 2% agarose gel at 120 V for 40 min (or as long as needed based on amplicon length) to confirm successful amplification (*see* **Note 5**).
8. Dilute first round PCR amplicons 1:100 by transferring 2 μl of PCR products to 98 μl H_2O in a new 96 well plate on ice. Seal the plate with an adhesive film, vortex, and centrifuge to bring all components to the bottom of the plate.

3.4 Barcoding

Second round PCR re-amplifies the first round amplicons using 5′ universal oligonucleotide anti-tags that are coupled to sample-specific barcodes. Ensure that all amplicons from the same genomic DNA sample are labeled with the same barcodes.

1. Place a 96 well PCR plate on ice for PCR preparation.
2. Prepare master mix for the reaction in a 1.7 ml tube. For each 20 μl reaction, add 16.25 μl H_2O, 2.5 μl 10× PCR buffer with Mg^{2+}, 4 μl dNTP (1.25 mM each), and 0.25 μl TaKaRa Taq HS.

3. Dispense 23 μl of master mix to each well.
4. Add 1 μl of universal primers with sample-specific barcodes (5 μM each) and 1 μl of diluted first round PCR products to each well.
5. Seal the plate with an adhesive film, vortex, and centrifuge to bring all components to the bottom of the plate.
6. Place the PCR plate in a thermal cycler and set the amplification program as following: initial denaturation step at 94 °C for 2 min followed by 35 amplification cycles (94 °C for 20 s, 60 °C for 45 s, and 65 °C for 1 min/kb + 30 s) and a final incubation at 65 °C for 5 min.
7. Electrophorese PCR products through a 2% agarose gel at 120 v for 40 min (or as long as needed based on amplicon length) to confirm successful amplification (*see* **Note 5**).

3.5 Amplicon Purification and Pooling

Prior to library construction and sequencing, amplicons require purification to ensure accurate quantification, which is critical to sample pooling and successful SMRT sequencing.

1. Purify PCR products using the QIAquick PCR purification kit according to the manufacturer's instructions (*see* **Note 6**).
2. Elute purified PCR products with 15 μl of elution buffer.
3. Measure concentration of purified PCR products (2 μl) with a NanoDrop 1000 and the dsDNA application.
4. Calculate the required volume of each amplicon using the following formula:

$$V_i = \frac{M \times L_i}{n \times C_{i \times \sum_{i=1}^{m} L_i}}$$

where V_i is the volume of each PCR amplicon, M is the total mass of pooled PCR amplicons (*see* **Note 7**), L_i is the length of each amplicon, n is the total number of samples, C_i is the concentration of each amplicon, and m is the total number of amplicons.

5. Add the calculated volumes of each amplicon into a new 1.7 ml tube. This pooled sample is now ready for SMRT sequencing library construction.

3.6 SMRT Sequencing Library Construction and Sequencing

SMRT sequencing libraries are constructed following the Pacific Biosciences Amplicon Template Preparation and Sequencing protocol, and SMRT sequencing performed according to the Pacific Biosciences P5-C3 protocol with a movie collection time of 180 min (*see* **Note 8**).

3.7 SMRT Sequencing Data Analysis

SMRT sequencing data analysis procedures were developed for users who prefer an independent graphical user interface (GUI)-based program, and for those who prefer a Linux environment bioinformatics pipeline.

We developed HiTMAP: a High Throughput Methylation Analysis Program to address the need for a stand-alone program capable of analyzing high-throughput targeted bisulfite sequencing data. HiTMAP takes raw, targeted bisulfite sequence data (FASTA) and demultiplexes against sample barcodes, aligns sequencing reads to in silico bisulfite-converted genomic reference sequences, quantitates CpG methylation levels, and exports resulting methylation data for both individual CpG sites and amplicon regions. The user-facing side of HiTMAP provides an online interface for uploading raw sequence and reference files, setting alignment, methylation quantitation, and quality metric parameters, and for retrieving and saving analysis output data and result figures (https://hitmap.stuartscottlab.org). Brief instructions on data submission to HiTMAP are noted below:

1. Upload FASTA files of SMRT-BS circular consensus sequence (CCS) reads to the HiTMAP homepage (*see* **Note 9**).
2. Upload reference sequence files for each of the targeted regions in FASTA format to the HiTMAP homepage.
3. Upload a barcode file for the sequenced samples in plain text CSV format to the HiTMAP homepage.
4. Click the "Submit" button for analysis.

The Linux environment data analysis pipeline, which implements Bismark [9] for SMRT-BS alignment and CpG methylation quantitation, is detailed below and illustrated in Fig. 2.

1. Using the FASTQ files of SMRT-BS CCS reads (*see* **Note 9**), demultiplex based on sample barcode using the NGSUtils tool kit [10] (*see* **Notes 10–12**):

```
fastqutils barcode_split \
  -edit 1 \
  -pos 20 \
  -allow-revcomp \
  -stats \
  <barcodes_file> \
  <input_fastq > \
  <output_template>
```

2. Remove short SMRT-BS reads from the data (*see* **Note 13**):

```
cat <fastq_file > |
awk '{if ((NR%4)==0) printf "%s\n",$0; else printf "%s\t",$0;}
'|
```

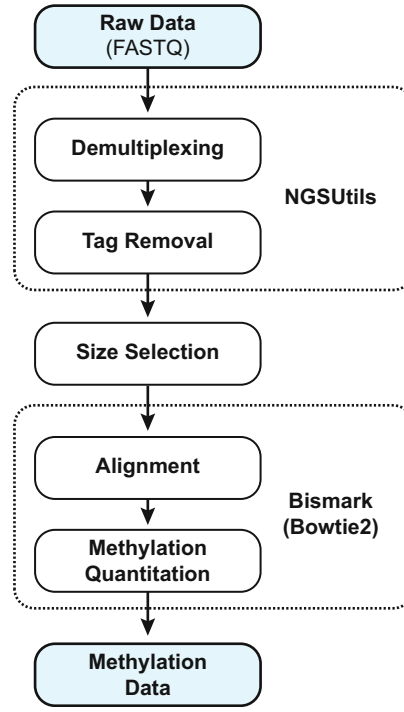


Fig. 2 Illustration of the SMRT-BS Linux-based analysis pipeline. Image adapted from Yang et al., *BMC Genomics*, 2015 [7]

```
awk '{if (length($2)>=200) print $0}' |
tr "\t" "\n" > <short_reads_removed.fastq>
```

3. Map SMRT-BS reads with Bismark, which invokes Bowtie2 [11] (*see* **Note 14**):

```
bismark \
–fastq \
–bowtie2 \
–non_directional \
-N 1 \
-L 5 \
-D 25 \
-R 3 \
-rdg 3,2 \
-rfg 3,2 \
–score_min L,0.6,-0.6 \
<Reference_file> \
<split_fastq>
```

4. Perform methylation quantitation with the Bismark methylation extractor:
   bismark_methylation_extractor \
   --single-end \
   --comprehensive \
   --report \
   --bedGraph \
   --counts \
   --cutoff 10 \
   --zero_based \
   --genome_folder directory_to_reference \
   <bismark_bt2.sam>

4 Notes

1. Several commercial bisulfite treatment kits were evaluated during the development of SMRT-BS, which identified two commercial kits that could enable PCR amplicons greater than 1 kb from bisulfite-converted DNA [7]; however, other commercial kits may be preferred based on individual user experience and experimental design.
2. Examining DNA by agarose gel electrophoresis prior to bisulfite conversion is recommended as degraded genomic DNA will lead to low yields of modified DNA.
3. Paired 18mer long barcodes for each sample is recommended so that SMRT-BS reads with only a single barcode (due to incomplete sequencing) can still be retrieved if desired. Ensure that barcodes are designed with regard to standard primer parameters such as GC content and sequence redundancy.
4. A lower extension temperature (e.g., 64–68 °C) than typical PCR (i.e., 72 °C) can lead to higher yields when amplifying A + T rich bisulfite converted DNA [7, 12].
5. Gel running times vary due to equipment used and length of PCR amplicons.
6. PCR product purification can also be done with Agencourt AMPure XP beads. When using AMPure XP beads, ensure that the appropriate volume of beads are used for amplicons with different lengths.
7. According to the Pacific Biosciences Amplicon Template Preparation and Sequencing protocol, the required input amount is 250 ng when amplicons are shorter than 750 bp, and 500 ng when amplicons are between 750 bp and 10 kb.

8. The P6-C4 chemistry from Pacific Biosciences could also be used for SMRT-BS. In addition, please note that SMRT-BS was developed on the PacBio RS II system; however, SMRT-BS could also be ran on the Sequel system.
9. Given that CpG methylation levels are calculated by read counting, it is recommended to use the CCS reads (as opposed to the less accurate subreads) for SMRT-BS data analysis.
10. For more information on the fastqutils barcode split tool, please refer to http://ngsutils.org/modules/fastqutils/barcode_split/.
11. If SMRT-BS reads are tagged with asymmetric barcodes, use the "cat" command in Linux to combine files for both barcodes after splitting.
12. FASTQ files from Pacific Biosciences SMRT sequencing have different quality encoding than the FASTQ files from Illumina sequencing. Please refer to the following post if fastq files are not recognized by the analysis tool.

    http://seqanswers.com/forums/showthread.php?t=48036.
13. Very short SMRT-BS reads can result from incomplete sequencing and/or sheared DNA fragments. As such, removing the extremely short reads (e.g., <200 bp when amplicons are ≥1000 bp) is recommended.
14. For more information on using Bismark, please refer to http://www.bioinformatics.babraham.ac.uk/projects/bismark/.

References

1. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A 89(5):1827–1831
2. Choufani S, Shapiro JS, Susiarjo M, Butcher DT, Grafodatskaya D, Lou Y, Ferreira JC, Pinto D, Scherer SW, Shaffer LG, Coullin P, Caniggia I, Beyene J, Slim R, Bartolomei MS, Weksberg R (2011) A novel approach identifies new differentially methylated regions (DMRs) associated with imprinted genes. Genome Res 21(3):465–476. doi:10.1101/gr.111922.110. doi:gr.111922.110 [pii]
3. Bliek J, Verde G, Callaway J, Maas SM, De Crescenzo A, Sparago A, Cerrato F, Russo S, Ferraiuolo S, Rinaldi MM, Fischetto R, Lalatta F, Giordano L, Ferrari P, Cubellis MV, Larizza L, Temple IK, Mannens MM, Mackay DJ, Riccio A (2009) Hypomethylation at multiple maternally methylated imprinted regions including PLAGL1 and GNAS loci in Beckwith-Wiedemann syndrome. Eur J Hum Genet 17(5):611–619. doi:10.1007/978-1-61779-612-8_17
4. Colyer HA, Armstrong RN, Sharpe DJ, Mills KI (2012) Detection and analysis of DNA methylation by pyrosequencing. Methods Mol Biol 863:281–292. doi:10.1007/978-1-61779-612-8_17
5. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A (2011) Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc 6(4):468–481. doi:10.1038/nprot.2010.190. doi:nprot.2010.190 [pii]
6. Adusumalli S, Mohd Omar MF, Soong R, Benoukraf T (2014) Methodological aspects of whole-genome bisulfite sequencing analysis. Brief Bioinform. doi:10.1093/bib/bbu016
7. Yang Y, Sebra R, Pullman BS, Qiao W, Peter I, Desnick RJ, Geyer CR, DeCoteau JF, Scott SA (2015) Quantitative and multiplexed DNA methylation analysis using long-read single-molecule real-time bisulfite sequencing

(SMRT-BS). BMC Genomics 16:350. doi:10.1186/s12864-015-1572-7. 1186/s12864–015–1572-7 [pii]

8. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S (2009) Real-time DNA sequencing from single polymerase molecules. Science 323 (5910):133–138. doi:10.1126/science.1162986. doi:1162986 [pii]
9. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27 (11):1571–1572. doi:10.1093/bioinformatics/btr167
10. Breese MR, Liu Y (2013) NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. Bioinformatics 29(4):494–496. doi:10.1093/bioinformatics/bts731
11. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9 (4):357–359. doi:10.1038/nmeth.1923
12. XZ S, Wu Y, Sifri CD, Wellems TE (1996) Reduced extension temperatures required for PCR amplification of extremely A+T-rich DNA. Nucleic Acids Res 24(8):1574–1575

Chapter 9

Copy Number Variation Analysis by Droplet Digital PCR

Suvi K. Härmälä, Robert Butcher, and Chrissy H. Roberts

Abstract

The health impact of many copy number variants in our genome remains still largely to be discovered. Detecting and genotyping this often complex variation presents a technical challenge. Here we describe a 96-well format droplet digital PCR (ddPCR) protocol for genotyping a common copy variant in the human haptoglobin gene. ddPCR allows for high-throughput and accurate quantitation of gene copy numbers.

Key words Droplet digital PCR, ddPCR, Copy number variation, CNV, Genotyping, Haptoglobin, HP

1 Introduction

Copy number variations (CNVs) account for a significant proportion of diversity in the human genome but the impact of CNVs on gene expression, protein function and disease traits is largely unknown [1, 2].

Detecting and enumerating CNVs using nucleic acid amplification techniques (NAATs) such as polymerase chain reaction (PCR) can be challenging. Quantitative PCR (qPCR) has previously been used to quantitate CNVs by comparing the dose–response of a CNV to a copy invariant gene (the reference) [3–6] but stringent optimization, calibration, quality control and a high number of technical replicates are required to obtain precise CNV estimates [4]. qPCR data are recorded in terms of the quantitation cycle (Cq) or cycle threshold (CT) value, which are both related to the starting concentration of the analyte on a $-\log_2$ scale. Here, the minus sign reflects that higher Cq/CT values indicate fewer initial copies of the analyte, whilst base 2 reflects the doubling of amplicon quantity that theoretically takes place with each additional PCR cycle. The Cq/CT difference between a specimen with one copy of the target and a specimen with two copies is just one cycle unit. The difference between one copy and three copies has a proportionally smaller interval at 1.6 cycle units. Weaver et al. [4] showed how the precise

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_9,

discrimination of one gene copy from two gene copies required around four technical replicates, whilst the discrimination of four gene copies from five gene copies would require upwards of 12 technical replicates [4]. Replicate qPCR tests often have coefficients of variation that approach plus or minus 1 cycle, with important sources of error coming from liquid handling, specimen complexity, inhibitory molecules and relative concentrations of the relevant analyte and irrelevant DNA moieties. All these sources of error can affect the amplification efficiency of PCR, thus invalidating the $-\log_2$ rule and leading to imprecise measures of copy number. These sources of error have the greatest magnitude effect at low copy numbers.

Digital PCR, an end-point PCR method, does not rely on assumptions of perfect amplification efficiency or on reference standards to reach high precision in copy number enumeration [7–9]. Through stochastic confinement and amplification of rare analytes in a plurality of nano-scale PCR reactors, digital PCR provides a direct molecular count of the analyte and reference target DNA sequences. This end-point PCR-based system is linear in nature, so in a diploid individual with four copies of the analyte per genome, the count data from the reference gene would be a number two times smaller than that for the analyte. Weaver et al. [4] showed that the precise discrimination of four gene copies from five gene copies was possible using a digital PCR system with upwards of 3080 partitions but at the time of that study high-throughput digital PCR systems were unavailable [4]. Droplet digital PCR (ddPCR, Bio-Rad Industries, Hemel Hempstead, UK) is one high-throughput implementation of digital PCR that is now widely available to the research community. During the ddPCR test, a PCR assay is partitioned into around 15,000 reverse micelles (i.e., water-in-oil) [8]. These droplets have a uniform 1 nL volume, each droplet is a PCR-competent nano reactor and the number of analyte molecules is calibrated so that it is substantially lower than the number of droplets [7, 8]. This process of stochastic confinement ensures that almost all of the droplets now contain either zero (the majority) or one copy (a minority) of the analyte sequence, with an increasingly small probability of containing two, three or more copies. This process happens in parallel for each of the analyte and reference sequences and the final pool of droplets constitutes two independent Poisson processes [10]. The droplets are thermostable and after undergoing PCR cycling in a normal thermal cycler, signals from fluorophore-linked hydrolysis probes indicate the presence of one, both or neither PCR target in a given droplet. The droplets are finally passed, in single file, through a "droplet reader", a modified flow cytometer with oil-based fluidics. The flow cytometer reads the fluorescence signals of each droplet and classifies the droplets as positive or negative for one or both of the target sequences [10] (Fig. 1). The number of positive droplets

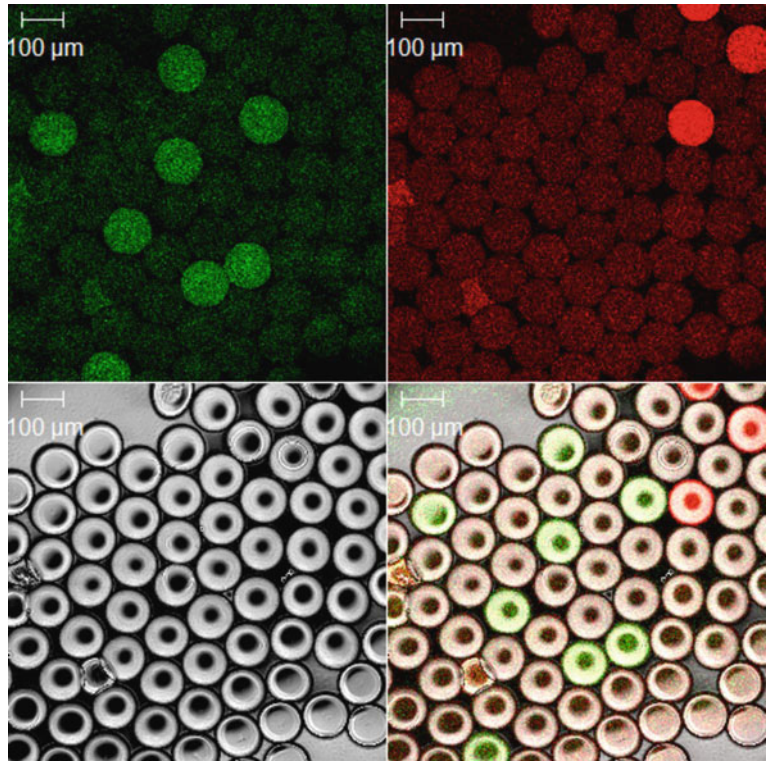


Fig. 1 Confocal photomicrograph of ddPCR droplets from a sample positive for both FAM and HEX targets post-ddPCR. A bright-field image of droplets is shown at the *bottom left*. Droplets positive for both the FAM channel (*green*; *top left*) and HEX channel (*red*; *top right*) are shown. A composite of the bright-field, FAM and HEX channels is shown at the *bottom right*. All droplets have noticeable baseline fluorescence on both channels. PCR-positive droplets fluoresce with much greater intensity than template-negative droplets. The majority of droplets are PCR negative. Figure taken from [11] and used under Creative Commons Attribution 3.0 Unported license (http:/creativecommons.org/licenses/by/3.0/)

for each assay is directly proportional to the starting number of analyte or reference molecules. The counts are converted to estimated concentrations through application of the Poisson calculation, thus accounting for the diminishing probabilities that any positive droplet contains 2, 3, 4, or more copies of the analyte. The ratio between the analyte and the reference concentrations indicates the copy number according to the formula $2 * (a/b)$ where 2 indicates that the sample is from a diploid organism with two genome copies of the reference gene, a is the concentration of the analyte and b is the concentration of the reference [10].

We developed a ddPCR assay to analyze the copy number variation in haptoglobin (*HP*). In the human body, haptoglobin protein binds to and chelates free hemoglobin that has been released from ruptured red blood cells [11]. The *HP* gene, located on chromosome 16 (16q22), has a 1.7 kb tandem two-exon segment copy number variant, also known as *HP1/2* [12]. As a consequence of this variation, three common genotypes, *HP1-1*, *HP1-2*,

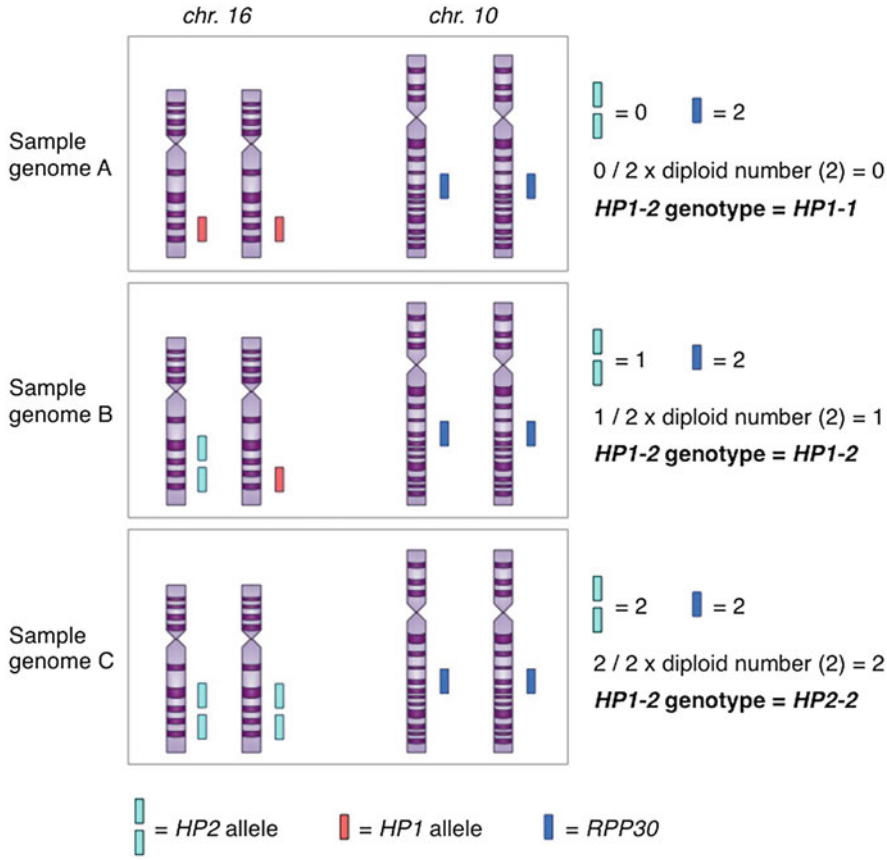


Fig. 2 ddPCR *HP1/2* genotyping assay principle. Number of *HP2* alleles present in each sample genome is compared to the number of copies of the reference gene *RPP30* in the same genome. Corrected for diploidy (multiplied by 2), this ratio determines the *HP1/2* genotype

and *HP2-2*, exist. Evidence suggests that this variation has a functional impact on the protein structure and binding properties [13–16]; however, the impact of this polymorphism on human health and disease is less clear. In our assay, we measured the number of *HP2* copies present in DNA by comparing the *HP2* DNA concentration to an unlinked, invariant gene on chromosome 10 (ribonuclease P protein subunit 30, *RPP30*) (Fig. 2). Here, we demonstrate the 96-well format protocol for genotyping *HP2* copy number variation in the human *HP* gene by ddPCR.

2 Materials

1. Molecular biology grade H_2O. Store at room temperature.
2. 1× TE buffer for molecular biology: 10 mM Tris–HCl, 1 mM disodium EDTA, pH 8.0. Store at room temperature.

Table 1
Oligonucleotides used in *HP* CNV genotyping

Oligonucleotide	Sequence
*HP*_forward	CCAGTGCTGCTCTAGATTCA (Fig. 3)
*HP*_reverse	GCACATCAATCTCCTTCCACC (Fig. 3)
*HP*_probe_FAM	FAM-GTAGCCCCTAGCCCTTTCAA-BHQ1 (Fig. 3)
*RPP30*_forward	AGATTTGGACCTGCGAGCG[a]
*RPP30*_reverse	GAGCGGCTGTCTCCACAAGT[a]
*RPP30*_probe_HEX	HEX-TTCTGACCTGAAGGCTCTGCGCG-BHQ1[a]

[a]Sequences taken from [22]

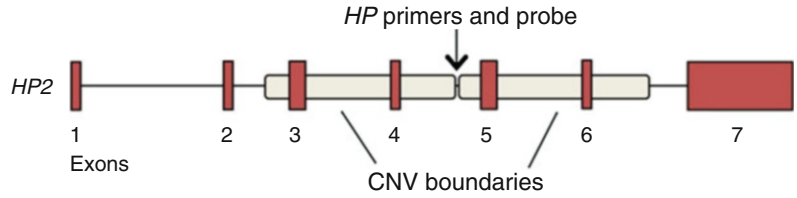


Fig. 3 Location of *HP* primers and probe on the *HP2* allele. Primers amplify a 165 bp sequence at the CNV breakpoint between exons 4 and 5

3. Purified DNA specimens (*see* **Note 1**).
4. Fluorometer and reagent kits for DNA quantification (*see* **Note 2**).
5. 100 μM primer/probe stock solutions: primers for *HP* target sequence and *RPP30* reference (Table 1, see also Fig. 3), 1× TE buffer (*see* **Notes 3** and **4**). Store at 4 °C.
6. 10× primer/probe mix: 100 μM primer/probe stock solutions, 1× TE buffer. Add 150 μL 1× TE buffer for molecular biology, 22.5 μL 100 μM *HP* forward primer stock solution, 22.5 μL 100 μM *HP* reverse primer stock solution, 22.5 μL 100 μM RPP forward primer stock solution, 22.5 μL 100 μM *RPP30* reverse primer stock solution, 5 μL 100 μM *HP* probe stock solution, and 5 μL 100 μM *RPP30* probe stock solution to a tube (total volume in the tube 250 μL) and mix thoroughly by vortexing and inverting the tube. Store at 4 °C.
7. 2× ddPCR supermix reagent (Bio-Rad) (*see* **Note 5**). Aliquot if needed and store for up to 48 h at 4 °C or at −20 °C for longer term storage. Avoid freeze–thawing.
8. Semi-skirted 96-well PCR plates.
9. Droplet generation cartridges and gaskets (both Bio-Rad).
10. Droplet generation oil (Bio-Rad). Store at room temperature.
11. Droplet generator (Bio-Rad).

12. Droplet generator cartridge holder (Bio-Rad).
13. Pierceable foil heat seals.
14. Heat sealer for plates.
15. Programmable thermal cycler.
16. Optical film compression pad.
17. Droplet reader oil (Bio-Rad).
18. Droplet reader (Bio-Rad).
19. QuantaSoft software (Bio-Rad).

3 Methods

The full protocol requires around 6–7 h. Total hands on time is typically around 1 h. Droplet generation takes approximately 30–40 min, PCR takes 1.5 h, and droplet reading takes 3 h. Data analysis takes 20–30 min per plate.

Based on 2016 list prices, the cost of this protocol is approximately GBP 3.37 per well, which does not include the cost of equipment, specimen preparation or operator time. The assay is highly sensitive, so it is important to take measures to avoid contamination of specimen or reagents. Prepare all reagent mixes (containing no DNA) in a dedicated PCR hood. Perform all steps at room temperature unless indicated otherwise.

3.1 Preparation of Specimen DNA

1. Estimate DNA concentration of each sample using fluorometer for DNA quantitation (*see* **Note 2**).
2. Dilute each sample to approximately 10 ng/μL in 1× TE for molecular biology or molecular biology grade H_2O (*see* **Note 6**). Incubate specimens for approximately 1 hr at room temperature to allow the concentration to equilibrate (*see* **Note 7**).

3.2 Preparation of PCR Reaction Mix

1. Prepare a PCR mix by adding 1100 μL of 2× ddPCR supermix reagent, 220 μL 10× *HP* primer/probe mix, and 220 μL 10× *RPP30* primer/probe mix into 660 μL molecular biology grade H_2O in a 5 mL tube (total volume 2200 μL prepared for 100 samples). Mix by repeatedly inverting and vortexing the tube to ensure that the viscous ddPCR supermix is thoroughly mixed with the less viscous components. Centrifuge briefly.
2. By reverse pipetting, add 21 μL of PCR mix into each well of a 96-well plate (*see* **Note 8**).
3. To complete the reaction mixes in the wells, carefully transfer 1 μL of each ~10 ng/μL DNA specimen (or control specimen) into the PCR mix aliquots on the 96-well plate, one specimen

Fig. 4 Droplet generation cartridge assembly with respective wells for input reaction mix, droplet generation oil and droplet output labeled. Cartridge wells following droplet generation, droplets illustrated as a cloudy suspension in the uppermost well (*inset*)

per well. Each reaction mix now contains 1× ddPCR supermix, 900 nM each primer, 200 nM each probe, and specimen DNA. At least one known-copy number positive and one known-negative well should be run on each plate to provide guidance for gating. Centrifuge plate at low speeds to collect reaction mix in the bottom of each well (Fig. 3).

3.3 Droplet Generation

1. Place a droplet generation cartridge in the cartridge holder (Fig. 4).
2. Using a multichannel pipette, carefully aspirate 22 μL of the reaction mix from the first column of the 96-well plate. Position the tips of the multichannel pipette at the bottom of the eight middle wells of the droplet generation cartridge (row labeled 'Sample') (Fig. 4). Expel the reaction mix. During expulsion, lift the tip upwards very slowly whilst always keeping the tip slightly below the surface of the liquid. It is important to not generate any air bubbles or foam in the lower part of the cartridge wells as this may cause the vacuum manifold to malfunction and prevent the droplets from being generated (*see* **Note 9**).
3. By reverse pipetting, add 60 μL of droplet generation oil to each cartridge wells in the row labeled "Oil" (Fig. 4).
4. Cover the cartridge carefully with a gasket and place the cartridge holder assembly in the droplet generator. Close the generator to start droplet generation.
5. When the droplet generation is complete, take the cartridge holder assembly out of the generator. Carefully remove the

gasket and observe the newly generated droplets in the row of wells marked "Droplets" of the cartridge as cloudy liquid (Fig. 4).

6. Set the multichannel pipette to 45 μL, position the tips just below the surface of the droplet solution, and gently lower it, following the liquid level as the droplet solution is aspirated. Aspiration should be very slow (~10 s) to avoid deforming the droplets. Transfer the droplets into the wells of the first column of a new 96-well semi-skirted plate. Expulsion of the droplets should be carried out with similar care to the aspiration step. Discard used cartridge and gasket and repeat the droplet generation steps for the remaining columns of the first 96-well plate. A total of 12 droplet generation cartridges, 12 gaskets, and 6000 μL (5750 μL plus excess) of droplet generation oil are needed for the droplet generation in this protocol.

3.4 PCR

1. Seal the droplet-containing 96-well plate with pierceable metallic heat seal in a plate sealer for 3 s at 170 °C. Turn the plate through 180° and repeat the sealing step to ensure the seal is secure.
2. Insert the sealed plate into the PCR machine and cover the plate with an optical film compression pad.
3. Run the following PCR program: 10′00″ at 95 °C, 40x (0′10″ at 95 °C, 0′30″ at 57 °C) 12′00″ at 98 °C, hold at 12 °C. The 12′00″ hold at 98 °C is a curing step in which the droplets gel to form semi-solid beads. When the PCR is complete, the plate can be used immediately for droplet reading, or can be stored at 4 °C for up to 48 h without significant loss of signal.

3.5 Reading Droplets

1. Ensure the droplet reader has sufficient droplet reading oil and that the waste container is not full.
2. Using the QuantaSoft control software, flush and prime the reader.
3. Open the droplet reader using the button on the reader and remove the metal retainer from the plate holder. Insert the PCR plate containing the droplets into the holder and replace and secure the retainer on top of the plate. Press the button on the reader to close the main cover.
4. On QuantaSoft software, each well must be assigned an experiment template, which should be customized for CNV experiments to reflect the expected copy number per cell of the reference gene. Select "Copy Number Variation" as the experiment type and number of reference copies as 2 for diploid genome.

5. Create a new plate template and highlight all the 96 wells on the plate template. In the "Sample detail" menu, select the experiment template you created. Designate FAM channel (Assay 1, blue) as "Unknown" (name it HP) and HEX channel (Assay 2, green) as "Reference" (name it RPP30). Enter the specimen identifiers in the individual cells of the table (*see* **Note 10**), click "Apply" to apply these settings to all wells and click "OK" to finish.
6. Click "Run" to begin reading droplets. Results of the run are saved automatically.

3.6 Analysis of Results

1. After the droplet reading is complete, manually determine threshold fluorescence values for FAM and HEX fluorophores in the 2D plot view of the software to determine the positive populations for each fluorophore in each sample. Where the target is present in a droplet, the fluorescence will be much higher than that of a droplet without a target (Fig. 1), and droplets will separate into populations based on this (*see* **Note 11**) (Fig. 5). Set these threshold fluorescence values conservatively just under main "body" of each of the higher fluorescence droplet populations (Fig. 5). Double check consistency of gating by rows of samples in a multi-sample 1D view and correct thresholds if necessary (Fig. 6).
2. Exclude any samples from further analysis that did not form clear population clusters or which had unusual patterns in droplet fluorescence, such as poor differentiation within or

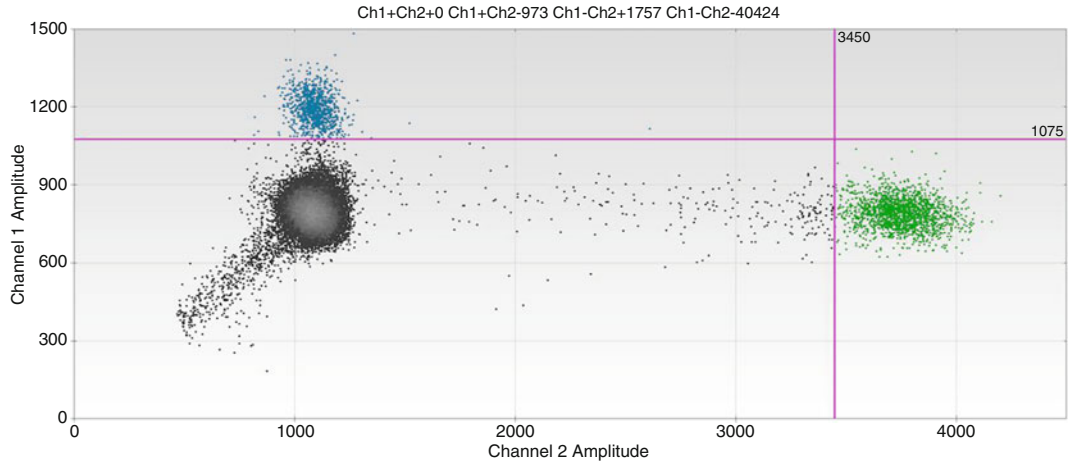


Fig. 5 QuantaSoft output with gates highlighted. In this 2D view, the droplet population in *blue*, above the set fluorescence threshold (*pink* line in intersecting the *y*-axis), is positive for FAM fluorophore (*HP2* target) (*see* **Note 12**). Droplet population in *green*, above the set fluorescence threshold (*pink* line intersecting the *x*-axis), is positive for HEX fluorophore (*RPP30* target). Droplet population in *grey*, below both fluorescence thresholds, consists of the droplets negative for both targets

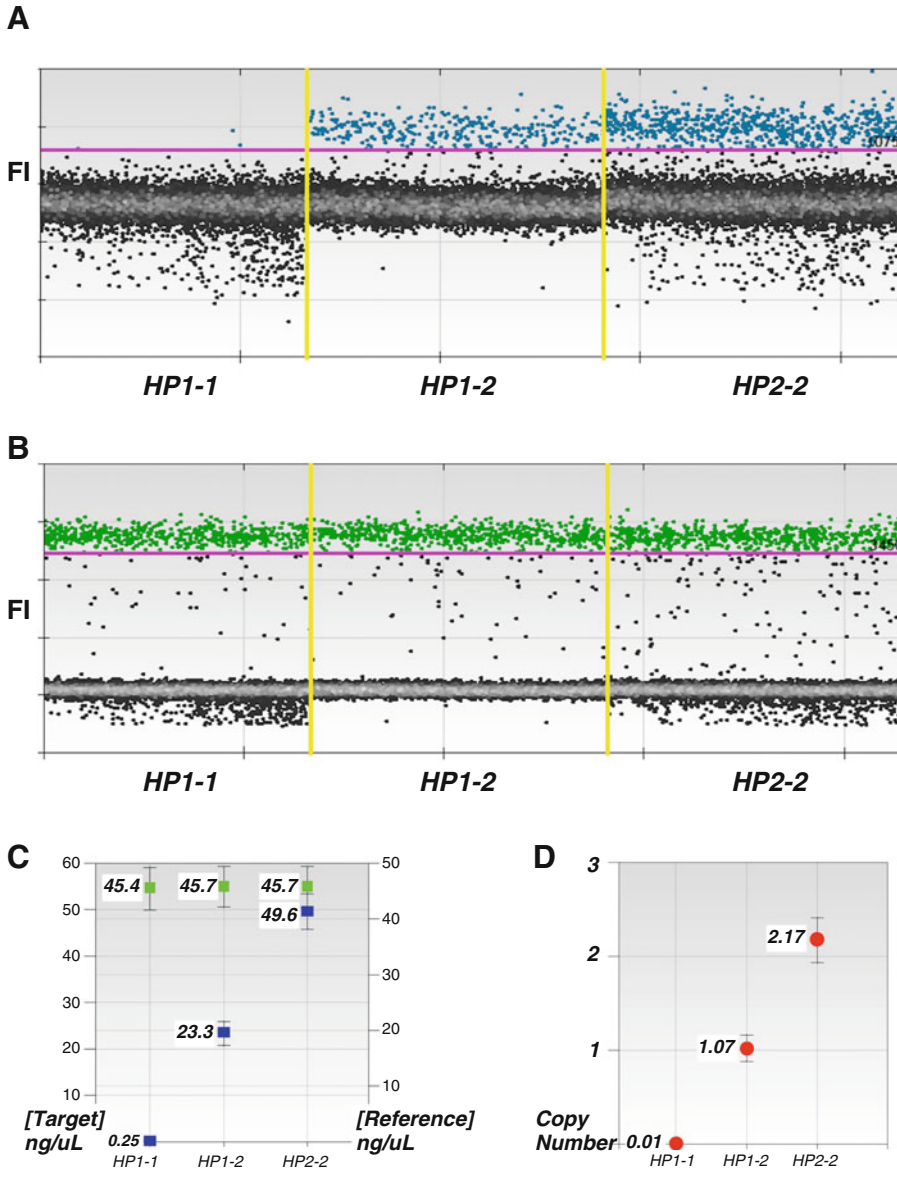


Fig. 6 Example of data output for each *HP* copy number variant *HP1–1*, *HP1–2*, and *HP2–2*. (**a**) Shows fluorescence intensity of FAM-positive (*HP2*-positive, in *blue*) and negative (*HP2*-negative, in *grey*) droplets. (**b**) Shows fluorescence intensity of HEX-positive (*RPP30*-positive, in *green*) and negative (*RPP30*-negative, in *grey*) droplets. (**c**) Shows the concentration of *HP2* target (copies/μL, *blue squares*) and *RPP30* reference (copies/μL, *green squares*) in the same specimens and (**d**) shows the copy number estimates (*red circles*)

between populations or more than two distinct populations for a single fluorescence channel.

3. Export results (except ones considered failed tests based on above) from the experiment to a Comma Separated Values (CSV) file in order to further manipulate the output in statistical software packages such as R [17].

Table 2
Suggested parameters of acceptability for copy number estimates

	Suggested acceptable parameters for copy number call	
Genotype	**Call**	**Confidence interval (CI)**
HP1–1	Call $\leq$ 0.05	0 $\leq$ CI $\leq$ 0.08
HP1–2	0.5 $\leq$ Call $\leq$ 1.2	0.5 $\leq$ CI $\leq$ 1.4
HP2–2	1.5 $\leq$ Call $\leq$ 2.6	1.5 $\leq$ CI $\leq$ 2.6

Estimates falling outside of these parameters should be discounted and re-run

4. We suggest tests should be rejected if the accepted droplet count is less than half the mean droplet count of the whole plate.
5. When there are no *HP2* alleles for the *HP* primers and probe to bind, no *HP* targets are detected in reading droplets and no copy number call is generated. In data exports this *HP1-1* genotype results in missing CNV data. To not exclude the samples with *HP1-1* genotype from the results, replace missing copy number call values with 0 (representing *HP1-1* genotype).
6. Copy number estimate can be influenced by many endogenous and exogenous factors such as DNA fragment size or purity and potentially DNA concentration. We suggest estimates outside the tolerances shown in Table 2 should be re-run (*see* **Note 13**).
7. Round the ddPCR copy number call of successful samples to the nearest integer.

4 Notes

1. For the workflow later on in this protocol, it might be easier to create a 96-well sample storage plate containing diluted samples ready to be assayed. Remember to include empty wells for positive and negative controls. Seal the plate when not in use. Where DNA samples are in a storage plate, briefly centrifuge the sealed plate before use of the specimen to ensure no sample droplets remain on the plate seal before unsealing.
2. We found quantification of highly concentrated, dense DNA samples to work better using fluorometric (we used Qubit Fluorometer and dsDNA BR Assay, ThermoFisher) compared to absorbance-based quantification.
3. Manufacturers often state in the technical datasheet or product specifications the μL amount of diluent to be added to achieve 100 μM primer stock. In case this information is not available,

it can be calculated based on the molecular weight and the gram weight of the primer using the formula below:

Weight in grams/molecular weight (g/mol)/target concentration M = volume to add in L.
Example:
Molecular weight of primer = 5711 g/mol.

Quantity in the vial = 363.2 μg = 0.0003632 g.

Target concentration for the stock = 100 μM = 0.0001 M.
Volume of 1× TE buffer for molecular biology to add to primer powder to achieve target concentration 100 μM = 0.0003632/5711/0.0001 = 0.000636 L = 636 μL.

4. The primers and internal probe targeting *HP2* allele-specific sequence were designed and screened for secondary structures using an online primer generation tool Primer3Web [18]. Additionally, the primers were screened for unwanted within assay interactions using free software AmplifX [19]. Presence of single nucleotide polymorphisms in the target sequence was controlled with the help of the NCBI Variation Viewer online tool [20]. Primer specificity was checked against human and other recorded genomes on the NCBI Nucleotide BLAST website [21]. The primers and internal probe targeting *RPP30* specific sequence were designed based on a previous study using this gene as their duplex test control gene [22]. Useful information about designing primers for ddPCR can be found in the droplet digital PCR application guide on the Bio-Rad website [23].
5. This protocol uses ddPCR system by Bio-Rad company. We recommend all equipment, materials, and reagents specific to droplet-based steps of this protocol are sourced from the same supplier.
6. Achieving exact concentration of 10 ng/μL can be tricky. Although the ideal concentration of DNA samples in this ddPCR assay is 10 ng/μL, we have performed this assay with DNA concentrations ranging from 7 to 25 ng/μL.
7. Note that this protocol only requires adjusting the concentration of the specimens as per above, it does not require any restriction enzyme treatment of the specimen. Some ddPCR copy number variation analysis protocols require physical separation of the variant copies to achieve an accurate copy number call [24]. In our assay, the *HP* target primers were designed to bind a sequence immediately prior to the repeat of the copy number variant two-exon segment, present only once on *HP2* allele (and absent from *HP1* allele). Consequently, our protocol does not require restriction endonuclease treatment to separate the two tandem copies of the *HP2* allele. No

physical separation was needed between *HP* and *RPP30*, as they are located far apart, on different chromosomes (chromosome 16 and 10, respectively).

8. Reverse pipetting reduces the risk of introducing air into the wells. To reverse pipette, push the pipette knob slightly beyond the first stop and place the tip into the PCR master mix. Slowly release the knob to rest position to aspirate in as much the pipette lets you. To release the PCR master mix into the target well, place the tip in the bottom of the well and slowly push the pipette knob to the first stop. Some PCR master mix remains in the tip. Instead of pushing this out, leave it in the tip, bring the tip back to the master mix tube and aspirate in another set of PCR master mix to replace the volume that was transferred into the well and transfer this volume into the next target well. To avoid contamination and wasting of precious reagents, transfer PCR mix into wells one well at a time with a single channel pipette. Do not use reagent reservoir and multichannel pipette.
9. Due to the structure of the droplet generation cartridge well bottoms, it might look like there is an air bubble in the middle of the cartridge well even when there isn't one. In case air bubbles do appear, they can be removed by touching the air bubble gently with a clean tip. Remember to change tips between samples.
10. Currently, a plate template cannot be uploaded into the software from an Excel or CSV file. We found easiest to name all samples in wells by writing plate number or name in the beginning of the sample name and then auto naming the wells by their plate location. For example, with all plate template wells highlighted, write Plate4_ on the sample name field, tick the Auto Inc. box to auto name and click Apply. Plate names will appear as Plate4_A07 for well A07 and Plate4_B01 for well B01 and so on. CSV files can be exported from the system and the temporary sample names can be easily matched and replaced with the real sample IDs using a program such as R [17].
11. Separation of positive and negative droplet populations can be improved by optimization of PCR parameters, such as oligonucleotide concentrations, annealing temperature or sample purity [25].
12. In a sample with *HP1-1* genotype, the *HP2* allele-specific target sequence is absent. As a result, PCR amplification does not occur, all droplets are *HP2*-negative and the FAM-positive (*HP2*-positive) droplet population seen in Fig. 5 will not be present in the output.
13. We noticed that often a sample with DNA concentration considerably higher than the suggested 10 ng/μL (such as

~25 ng/μL) resulted in "in between" copy number calls (such as 0.3 or 1.3). Samples with lower than suggested DNA concentrations (such as ~7 ng/μL), on the other hand, were observed to result in wide copy number call confidence intervals. It might be worth checking the DNA concentrations from the ddPCR results for those samples that fail to achieve a copy number call within the suggested parameters (Table 2), re-dilute as necessary and re-run these samples with the adjusted DNA concentration. Given that sample DNA was diluted 1:22 in the PCR reaction mix, one copy of haploid genome = 3 pg and 1000 pg = 1 ng, DNA concentration of the assayed samples can be calculated from the ddPCR results as follows: DNA concentration (ng/μL) = *RPP30* concentration from droplet reader (copies/μL) * 22 * 3/1000.

Acknowledgments

This work was supported by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement (MC-A760-5QX00). Further funding was provided by the UK Biotechnology and Biological Sciences Research Council (BBSRC BB/M009513/1 to SKH). We thank BJ Hennig for access MRC Keneba Biobank (The Gambia) samples and data, and special thanks to all participants and staff at MRC Keneba, The Gambia. RB is supported by the Wellcome Trust (098521/B/12/Z). ChR is supported by the Wellcome Trust Institutional Strategic Support Fund (105609/Z/14/Z).

References

1. Henrichsen CN, Chaignat E, Reymond A (2009) Copy number variants, diseases and gene expression. Hum Mol Genet 18:1–8
2. Buchanan JA, Scherer SW (2008) Contemplating effects of genomic structural variation. Genet Med 10:639–647
3. Jiang W, Johnson C, Jayaraman J et al (2012) Copy number variation leads to considerable diversity for B but not a haplotypes of the human KIR genes encoding NK cell receptors. Genome Res 22:1845–1854
4. Weaver S, Dube S, Mir A et al (2010) Taking qPCR to a higher level: analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. Methods 50:271–276
5. D'haene B, Vandesompele J, Hellemans J (2010) Accurate and objective copy number profiling using real-time quantitative PCR. Methods 50:262–270
6. Rose-Zerilli MJ, Barton SJ, Henderson AJ et al (2009) Copy-number variation genotyping of GSTT1 and GSTM1 gene deletions by real-time PCR. Clin Chem 55:1680–1685
7. Hindson BJ, Ness KD, Masquelier DA et al (2011) High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal Chem 83:8604–8610
8. Pinheiro LB, Coleman VA, Hindson CM et al (2012) Evaluation of a droplet digital polymerase chain reaction format for DNA copy number quantification. Anal Chem 84:1003–1011
9. Hindson CM, Chevillet JR, Briggs HA et al (2013) Absolute quantification by droplet digital PCR versus analog real-time PCR. Nat

Methods 10(10):1003–1005. doi:10.1038/nmeth.2633
10. Roberts CH, Jiang W, Jayaraman J et al (2014) Killer-cell immunoglobulin-like receptor gene linkage and copy number variation analysis by droplet digital PCR. Genome Med 6:20
11. Langlois MR, Delanghe JR (1996) Biological and clinical significance of haptoglobin polymorphism in humans. Clin Chem 42:1589–1600
12. Maeda N, Yang F, Barnett DR et al (1984) Duplication within the haptoglobin Hp2 gene. Nature 309:131–135
13. Kristiansen M, Graversen JH, Jacobsen C et al (2001) Identification of the haemoglobin scavenger receptor. Nature 409:198–201
14. Nielsen MJ, Moestrup SK (2009) Receptor targeting of hemoglobin mediated by the haptoglobins: roles beyond heme scavenging. Blood 114:764–771
15. Okazaki T, Yanagisawa Y, Nagai T (1997) Analysis of the affinity of each haptoglobin polymer for hemoglobin by two-dimensional affinity electrophoresis. Clin Chim Acta 258:137–144
16. Wejman JC, Hovsepian D, Wall JS et al (1984) Structure and assembly of Haptoglobin polymers by electron microscopy. J Mol Biol 174:343–368
17. R Core Team (2015) R: a language and environment for statistical computing. In: R Found. Stat. Comput. https://www.r-project.org/. Accessed 31 Jul 2016
18. Rozen S, Skaletzky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics methods and protocols (methods in molecular biology). Humana Press, Totowa, NJ, pp 365–386
19. Jullien N (2013) AmplifX 1.7.0. http://crn2m.univ-mrs.fr/pub/amplifx-dist. Accessed 31 Jul 2016
20. National Centre for Biotechnology Information (2016) Variation Viewer http://www.ncbi.nlm.nih.gov/variation/view/. Accessed 31 Jul 2016
21. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410
22. Luo W, Yang H, Rathbun K et al (2005) Detection of human immunodeficiency virus type 1 DNA in dried blood spots by a duplex real-time PCR assay detection of human immunodeficiency virus type 1 DNA in dried blood spots by a duplex real-time PCR Assay. J Clin Microbiol 43:1851–1857
23. Bio-Rad (2014) Droplet digital PCR applications guide. pp. 1–111. http://www.bio-rad.com/webroot/web/pdf/lsr/literature/Bulletin_6407.pdf/. Accessed 5 July 2017
24. Karlin-Neumann G, Montesclaros L, Heredia N et al (2012) Probing copy number variations using bio-Rad's QX100™ droplet digital™ PCR system. BioRad Tech Bull 6277. http://www.bio-rad.com/webroot/web/pdf/lsr/literature/bulletin_6277.pdf. Accessed 5 July 2017
25. Dingle TC, Hall Sedlak R, Cook L, Jerome KR (2013) Tolerance of droplet-digital PCR versus real-time quantitative PCR to inhibitory substances. Clin Chem 59:1670–1672

Chapter 10

MicroScale Thermophoresis: A Rapid and Precise Method to Quantify Protein–Nucleic Acid Interactions in Solution

Adrian Michael Mueller, Dennis Breitsprecher, Stefan Duhr, Philipp Baaske, Thomas Schubert, and Gernot Längst

Abstract

Interactions between nucleic acids and proteins are driving gene expression programs and regulating the development of organisms. The binding affinities of transcription factors to their target sites are essential parameters to reveal their binding site occupancy and function in vivo. Microscale Thermophoresis (MST) is a rapid and precise method allowing for quantitative analysis of molecular interactions in solution on a microliter scale. The technique is based on the movement of molecules in temperature gradients, which is referred to as thermophoresis, and depends on molecule size, charge, and hydration shell. Since at least one of these parameters is typically affected upon binding of a ligand, the method can be used to analyze any kind of biomolecular interaction. This section provides a detailed protocol describing the analysis of DNA–protein interactions, using the transcription factor TTF-I as a model protein that recognizes a 10 bp long sequence motif.

Key words Binding assay, Dissociation constant, DNA–protein interactions, MicroScale thermophoresis, Binding affinity

1 Introduction

The genes that code for the eukaryotic ribosomal RNAs, forming the major structural and functional part of the ribosome, are located in the nucleolus and are transcribed by RNA polymerase I. About 65–75% of total cellular transcription is exerted by RNA polymerase I and it is the only cellular polymerase that requires a specific DNA binding protein for transcription termination. There are multiple terminator elements located downstream of the gene and recognized by a 130 kD protein, termed Transcription termination factor I (TTF-I) [1]. Transcription of murine rRNA gene terminates downstream of the 3′ end of 28S RNA, involving the interaction of TTF-I with the repeated terminator elements [1]. TTF-I exhibits a modular structure, consisting of a C-terminal DNA-binding domain and a central domain that is required for

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_10,

transcription termination [2]. Besides the role of TTF-I in transcription termination, it is a chromatin-specific factor that also binds to the rRNA gene promoter, inducing changes of the chromatin structure and thereby enabling gene activation [3, 4].

Essential for understanding the function of TTF-I and of other DNA binding factors are quantitative parameters such as DNA binding site affinities. **M**icro**S**cale **T**hermophoresis (MST) represents a powerful and well-suited technology to quantify the affinities of protein–nucleic acid interactions.

MST is based on the directed movement of molecules along temperature gradients, an effect termed thermophoresis [5–7]. A spatial temperature difference ΔT leads to depletion of molecule concentration in the region of elevated temperature, quantified by the Soret coefficient S_T: $c_{\text{hot}}/c_{\text{cold}} = \exp(-S_T \Delta T)$. The directed movement of molecules through temperature gradients is depending on their size, charge, and hydration shell. Upon binding of a ligand to a molecule, at least one of these parameters is changed, resulting in distinct thermophoretic movements of the unbound and bound states [8].

As shown in Fig. 1a, the MicroScale Thermophoresis technology uses optics to monitor the thermophoresis of molecules through temperature gradients by detecting the optically visible molecule in aqueous buffer, in capillaries with a volume of 4–6 μL. Either intrinsically fluorescent molecules such as proteins via tryptophane residues [9] or molecules with an attached fluorophore can be used [10, 11]. An infrared laser establishes a microscopic temperature difference ΔT, which spans 2–6 °C, depending on the instrument settings.

Figure 1b represents a typical MST experiment. In the initial 5 s of the experiment, homogeneity is tested by monitoring the fluorescence in the sample in the absence of the temperature gradient. In the following, the temperature gradient is established by an IR laser. This causes an initial steep drop of the fluorescence signal—the so-called Temperature- or *T*-Jump—which reflects the temperature dependence of the fluorophore quantum yield. After the *T*-Jump, a slower thermophoresis-driven depletion of fluorophores occurs. Upon deactivation of the laser, a reverse *T*-Jump and concomitant back-diffusion of fluorescent molecules can be observed. In case of thermophoresis, movement of molecules from hot to cold regions is referred to as positive thermophoresis, while movement from cold to hot regions is referred to as negative thermophoresis.

Binding parameters of a molecular interaction can be determined by MST since thermophoresis correlates with molecular properties such as size, hydration shell, and charge, which are typically altered upon binding of a ligand. In a MST experiment, a serial dilution of the ligand is prepared and mixed with a constant concentration of labeled target molecule to establish different ratios

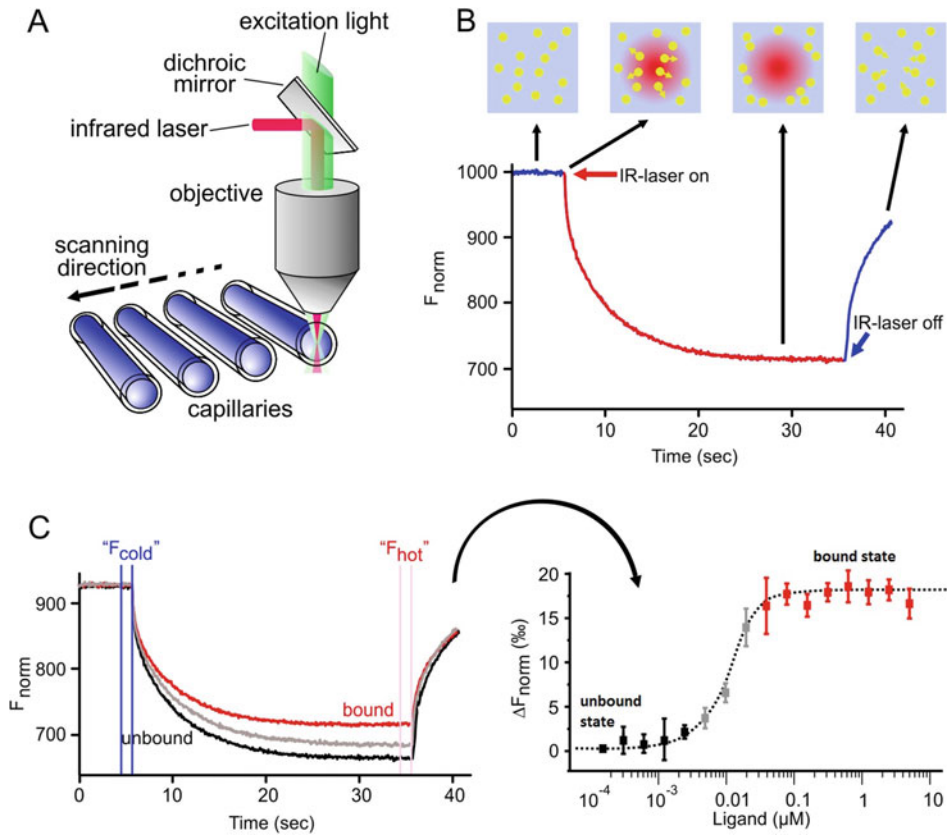


Fig. 1 (**a**) The technical setup of the MST technology is shown. Optics focus in the center of glass capillaries, thereby detecting the fluorescence signal of the optical visible molecule. An IR laser is used to establish a temperature gradient in the observation window of the optical system. Changes in fluorescence can be used to monitor thermophoretic movement of the molecules in solution (**b**) MST time trace—movement profile of molecules in a temperature gradient. After an initial 5 s cold phase (*laser off*), the laser is switched on and establishes the temperature gradient. After the *T*-Jump phase, in which the fluorescent dye decreases its signal yield upon heat induction, the thermophoretic movement takes place. After 30 s the laser is turned off and the molecules diffuse back. (**c**) Results of a typical MST experiment: The MST time traces of 16 capillaries containing the same concentration of optically visible interaction partner and an increasing concentration of the unlabeled ligand are recorded and plotted on one graph (*left*). The normalized fluorescence of the MST traces is plotted against the concentration of the ligand (*right*). The data points are fitted to obtain binding parameters such as the binding affinity

of the binding partners. The samples are loaded into capillaries and analyzed in the instrument by subsequently scanning each capillary. The resulting movement profiles (MST time traces) of the different ratios of interaction partners are plotted in one graph (Fig. 1c, left). Quantitative information on binding parameters can be extracted from the data (Fig. 1c, right).

This section provides a protocol to analyze the binding affinity of the transcription factor TTF-I to different DNA sequences: In the described experiments the different Cy5-labeled dsDNAs are

kept constant in each capillary whereas the concentration of TTF-I is stepwise reduced by 16 sequential 1:1 serial dilutions. Changes in DNA thermophoresis upon TTF-I binding are measured and plotted as a function of the protein concentration, and the dissociation constant is calculated.

The experimental scheme outlined below can be adapted for most protein–nucleic acid binding measurements.

2 Materials

2.1 Buffers and Reaction Partners

2.1.1 Reaction Buffer

MST buffer.

50 mM Tris–HCl pH 7.8, 150 mM NaCl, 10 mM $MgCl_2$, 0.05% Tween-20.

2.1.2 Preparation of the DNA Template

1. Fluorescently labeled oligonucleotides (Metabion) were ordered with the following sequences: T1 binding site of TTF-I: 5′-Cy5- CTT CGG AGG TCG ACC AGT ACT CCG GGC GAC-3′ and the complementary strand 5′- GTC GCC CGG AGT ACT GGT CGA CCT CCG AAG-3′. Control DNA: Cy5–5′-TCT TTT TTT TTT TTT TTC TTT TTT CCT CCT TTT TTT TTC C-3′ and the complementary strand: 5′- GGA AAA AAA AAG GAG GAA AAA AGA AAA AAA AAA AAA AAG A-3′. Once dissolved in water oligonucleotides were stored in light protected vials at −20 °C.
2. Annealing buffer (10×): 10 mM Tris–HCl pH 8.0, 50 mM NaCl, 1 mM EDTA. Stored at room temperature.
3. TBE buffer (0.4×): 35 M Tris, 35 mM Boric acid, 0.8 mM EDTA pH 8.0.
4. 30% Acrylamide/bis-acrylamide solution (37.5:1, Roth). Avoid unnecessary exposures, as the unpolymerized solution is neurotoxic.
5. *N*,*N*,*N*′,*N*′-Tetramethyl-ethylenediamine (99% p.a.) for electrophoresis (TEMED, Roth). Stored at 4 °C.
6. Ammonium persulfate (APS): prepare 20% solution in water and store aliquots at −20 °C.
7. Gel chamber system such as XCell Sure Lock™ System (Invitrogen).
8. GeneRuler™ Ultra Low Range DNA Ladder (Fermentas).
9. Glycerol >99.5% p.a. (Roth).
10. Ethidium bromide (Roth) stored at room temperature and a dark place: Prepare a fresh 1:10,000 solution in water before use. Beware that the chemical is toxic and mutagenic, so avoid contact and wear adequate protection.

11. FLA-5100 Fluorescence Imager (Fujifilm).
12. UV/VIS Spectrophotometer such as Nanodrop (Peqlab).

2.1.3 TTF-I Protein

TTF-I protein was expressed in *E. coli* and purified via its His-tag. The elution buffer was replaced by the storage buffer Ex100 (100 mM KCl, 20 mM Tris–HCl pH 7.6, 1.5 mM $MgCl_2$, 0.5 mM EDTA pH 8.0, 10% Glycerin) before freezing the protein in liquid N_2 and storage at −80 °C. The stock concentration of TTF-I was 442 micro M according to Bradford measurements.

2.2 Microscale Thermophoresis Equipment

1. Microscale Thermophoresis instrument Monolith NT.115 (NanoTemper Technologies, Munich, Germany).
2. Monolith NT™ capillaries purchased from NanoTemper technologies (Standard treated, Hydrophobic or Premium).

3 Methods

3.1 Annealing of Oligonucleotides and Preparation of DNA Working Solution

Double-stranded DNA substrate molecules were annealed from single-stranded oligonucleotides. It is crucial for the experiment that the fluorescently labeled oligonucleotide is quantitatively incorporated into the DNA substrate. This is achieved by adding the unlabeled oligonucleotide at a 1.15-fold molar ratio with respect to the labeled oligonucleotide to the annealing reaction. The efficiency of the annealing reaction can be determined on a 15% native polyacrylamide (PAA) gel that is first analyzed on a fluorescence imager to reveal non-incorporated, fluorescently labeled oligonucleotides and second, post-stained with ethidium bromide.

1. Dissolve oligonucleotides according to the manufacturer's instructions in water and measure the nucleic acid concentration using a UV/VIS Spectrophotometer.
2. Mix 575 pmol unlabeled oligonucleotides with 500 pmol Cy5-labeled oligonucleotide. Then, add 5 μL annealing buffer (10×) and adjust the volume to 50 μL with ddH_2O to finally obtain a 10 μM solution of double-stranded DNA.
3. Incubate the mixture for 5 min at 95 °C on a thermoblock, then switch off the thermoblock and allow the reaction to slowly cool down until it reaches room temperature. The reaction can now be stored at −20 °C.
4. A 15% native PAA gel is prepared by the following scheme and quickly poured into an assembled gel chamber: 9 mL 30% acrylamide/bis-acrylamide, 9 mL 0.4× TBE, 25 μL APS, 5 μL TEMED. Position a ten-well comb in the top of the gel. After the gel is polymerized (60 min), place the chamber into the running cell, remove the comb, and fill it with 0.4× TBE

running buffer. To remove unpolymerized acrylamide, pre-run the gel for 30 min at 100 V.

5. 150 nmol of the annealing reaction, as well as 150 nmol of the single-stranded oligonucleotides are individually mixed with glycerol to reach a final concentration of 5% (v/v) glycerol. This will weigh down the sample and prevent the solution to mix with the buffer in the well. Load carefully all samples together with the DNA ladder onto the pre-run gel, connect it to a power supply, and run it at 4 °C for 90 min at 120 V. Bromophenol Blue (usually present in the DNA marker) can be used as an indicator, as it migrates ahead of the single-stranded oligonucleotides with an apparent molecular weight corresponding to an oligonucleotide of about 10 nt in length.
6. The gel is visualized with a fluorescence imager. The fuzzy bands of the individual oligonucleotides have to be quantitatively shifted to a higher migrating, sharp band in the annealing reaction, representing the hybridized oligonucleotides.
7. Optionally, the efficiency of the annealing reaction is monitored by ethidium bromide staining. The gel is placed in the aqueous ethidium bromide solution and shaken for 10 min at room temperature. The gel can subsequently be visualized on a UV screen.
8. If free, labeled oligonucleotides are visible, the annealing reaction has to be repeated, increasing the ratio of the fluorescently labeled oligonucleotide.
9. Prepare 400 μL of the DNA working solution in MST buffer with a concentration of 62.5 nM. Please note, that the final concentration of DNA per capillary will be 50 nM.

General notes: To determine dissociation constants from a serial dilution, the concentration of the fluorescently labeled molecule should be close to or below the expected K_d. For optimal results, the concentration of the fluorescently labeled molecule and the LED power of the Monolith NT.115 instrument should be adjusted in such a way that the observed fluorescence intensity lies between 200 and 1500 fluorescence counts. Low excitation intensities (low LED powers) are suggested to reduce photobleaching effects (*see* **Note 1**).

3.2 Preparation of the Titration Series

A titration series consists of up to 16 capillaries which are measured in a single thermophoresis run. Dilutions of the unlabeled TTF-I should start at a concentration at least about 40-fold higher than the expected K_d. Notice that pipetting the samples and filling the capillaries will take about 30 min in total.

1. Prepare a 16-step 1:1 (v/v) serial dilution of TTF-I stock (442 μM) in the reaction buffer, so that each dilution step

reduces the protein concentration by 50%. For this, 16 small micro reaction tubes (200 µL) should be prepared: Label the reaction tubes from 1 through 16. Fill 12 µL of TTF-I in tube 1. Now add 6 µL of reaction buffer into the micro reaction tubes 2–16. Transfer 6 µL of tube 1 to tube 2 and mix thoroughly by pipetting up and down several times, transfer 6 µL to the next tube and repeat this dilution for the remaining tubes. It is important to avoid any buffer dilution effects. The buffer in tube 1 and the buffer in the tubes 2–16 must be identical.

2. BSA is included in this study as control protein for the specificity of TTF-I binding. Therefore, prepare a serial dilution with BSA which is comparable to that of TTF-I.

Please note that the NanoTemper analysis software contains a function (concentration finder) that can be used to determine the optimal concentration range of the titration partner (*see* **Note 2**).

3.3 Preparation of the Final Reaction mix

For the ease of pipetting and the minimization of experimental errors, the individual binding reactions should be prepared with an optimal volume of 30 µL (24 µL DNA working solution +6 µL of the respective TTF-I dilution). However, a volume of only 6 µL is sufficient to fill the capillary.

1. Add 24 µL of the 62.5 nM DNA working solution to 6 µL of each TTF-I dilution (or BSA dilution) (*see* Subheading 2 of Chapter 3). Mix the sample by pipetting and briefly centrifuge the samples. Consider this initial dilution step when calculating the final concentrations of TTF-I and DNA.
2. Incubate the samples for 5 min and fill the samples into standard capillaries.

For information on how to choose the right type of capillary (*see* **Note 3**).

Please note, powder-free gloves should be used to fill capillaries, thereby preventing impurities and adverse effects on the glass surface. In addition, the capillaries should not be touched in the middle where the measurement is performed.

Capillaries are filled by dipping into the samples. Since adhering molecules may falsify the measurement, care should be taken that the capillary does not touch the surface of the reaction tube. The capillaries are placed onto the capillary holder tray, which is then inserted into the Monolith NT.115 instrument. Use the NT Control software to set up experiment parameters and start the MST experiment (*see* below).

3.4 Capillary Scan

Besides information on the position of capillaries in the tray, the capillary scan provides important information about sample quality. Therefore, this scan is performed prior to the MST measurement to

detect adsorption of fluorescent molecules to the capillary walls (*see* **Note 3**), and pipetting errors or fluorescence quenching effects. Hence, the capillary scan is an important quality control to quickly identify irregularities and to accordingly optimize the glass capillary type, buffer conditions, or sample quality.

1. After starting the NT Control software, select the "red" LED channel for Cy5-dyes. Press the button "start capillary scan" to initiate the capillary scan (initial settings: LED Power: 20%).

The fluorescence signal during the capillary scan should be between 200 and 1500 fluorescence units (Monolith NT.115). If the value is below 200 fluorescence units, please refer to **Note 4**. Since all samples should contain the same concentration of fluorescently labeled DNA, individual differences in intensity between capillaries should be below 10% (*see* **Note 5** for trouble shooting if the variance in overall fluorescence is larger).

3.5 MST Measurement

After completing the capillary scan, the MST measurement can be started.

1. Initially, the TTF-I concentrations from the dilution series to the respective capillary position have to be assigned. For this, enter the highest concentration of TTF-I (final 88,400 nM) for capillary 1, select the correct dilution type (in this case 1:1), click on the maximum concentration, and use the implemented drag-function to automatically add the remaining concentrations for capillaries 2–16.
2. Enter the final concentration of fluorescent DNA (50 nM).
3. Next the desired reaction temperature has to be selected. Enable and activate the temperature control and select 25 °C for this experiment. Please note that NanoTemper Monolith™ devices are temperature controlled in a range from 20 to 45 °C.
4. In the following the laser power has to be adjusted. As previous experiments showed an optimal MST signal at 80% laser power, these conditions are chosen. Keep in mind that the strength of the temperature gradient induced by the IR laser correlates with the MST power. Typically, an experiment is started with lower MST power (20%). For more information on the MST power, *see* **Note 6**.
5. When using default settings, the fluorescence is initially detected for 5 s without temperature gradient. Upon switching on the laser, the thermophoresis is recorded for 30 s. After inactivation of the laser, the fluorescence is recorded for additional 5 s, monitoring back-diffusion of molecules. For high MST powers (>80%), we recommend reducing the "MST ON" time to <15 s.

6. In order to save the experimental data select a destination folder and create a data file. The MST measurement is started by clicking the start button. Using the abovementioned settings, one measurement will be completed within 10–15 min.
7. Repeat the measurements two additional times for a more accurate determination of the K_d value. In general, biological repeats are recommended.

3.6 MST Data Analysis

The integrated NT-Analysis software allows data analysis already during data acquisition. A plot of typical MST traces and a plot of the changes in the normalized fluorescence (F_{norm}) versus the ligand concentration are shown in Fig. 1b, c, left. Both plots are important for data analysis. The MST time traces offer important information on aggregation and precipitation effects, thus representing another important quality control feature of the MST technology (*see* **Note 7**).

In order to obtain binding constants from MST traces, the ligand-dependent changes in normalized fluorescence F_{norm} are calculated with

$$F_{norm} = F_{(hot)}/F_{(cold)},$$

where $F_{(hot)}$ and $F_{(cold)}$ represent averaged fluorescence intensities at defined time points of the MST traces. By default, three different settings can be chosen to analyze the data: Thermophoresis, Thermophoresis + *T*-Jump, and *T*-Jump only. For information on the different settings, *see* **Note 8**. Once F_{norm} for the chosen cursor settings is plotted, the data can be fitted to obtain either the dissociation constant (K_d) or the EC50 value (Fig. 1c, right). For more information on the curve fitting formula, *see* **Note 8**.

Step-by-step data analysis:

1. Press the "load project" button to import the acquired data, and select the data set and the MST runs you want to analyze.
2. Choose either the Thermophoresis, Thermophoresis and *T*-jump, or *T*-jump cursor settings. The respective F_{norm} values will be plotted. For the analysis of TTF-I-DNA data, select the *T*-Jump and Thermophoresis cursor settings.
3. The dissociation constant K_d of the interaction is determined by fitting the data using MST-standard fit algorithms (law of mass action) (Fig. 2a). F_{norm} values and the corresponding fit can be exported as a text or excel file, and results can also be summarized as a report in pdf format.
4. A better side-by-side comparison can be achieved by normalization (Fig. 2b) to the fraction of complexed molecules (FB) by the following equation:

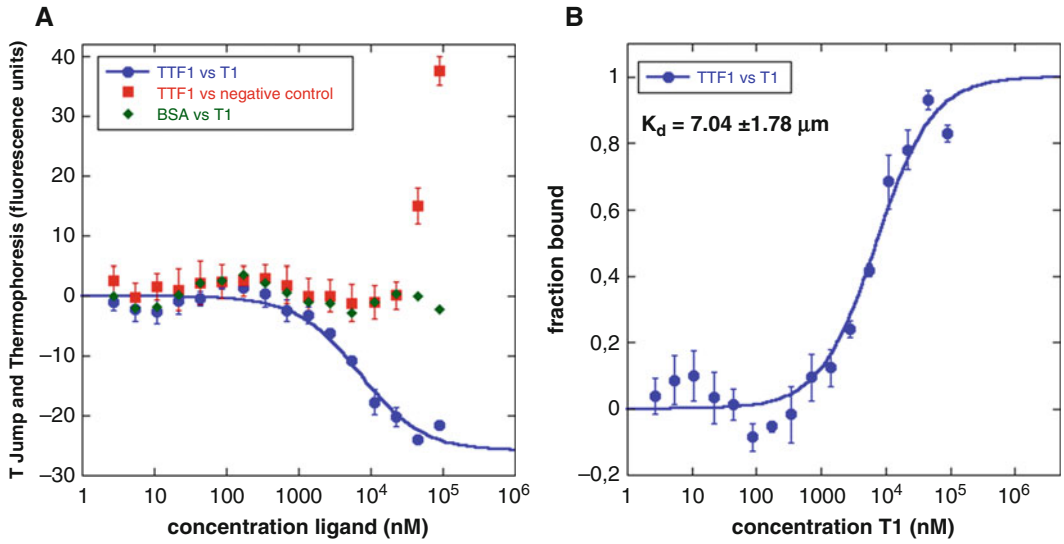


Fig. 2 MST data analysis. (**a**) Plot of the normalized fluorescence F_{norm} (‰) from *T*-Jump and Thermophoresis vs. the concentration of ligand (TTF-I or BSA). Lines represent fits of the data points using the K_d fit derived from law of mass action. STDEV derives from three biological repeats. (**b**) Fraction bound plot of the TTF-I vs. T1 data shown in Fig. 3a. STDEV derives from three biological repeats

$$\mathrm{FB} = (value(c) - free)/(complexed - free),$$

where $value(c)$ is the MST value measured for the concentration c, *free* is the MST value for the unbound state (lowest concentration), and *complexed* is the MST value for the fully bound state.

In this study, the interaction of the transcription factor TTF-I with its binding site T1 and an unspecific DNA template was determined by MicroScale Thermophoresis. TTF-I was shown to bind to its specific binding site T1 with an affinity of 7.04 ± 1.78 µM. BSA showed no binding to DNA and TTF-I did not interact with a control DNA of similar length. These results demonstrate that MicroScale Thermophoresis is an easy-to-use, fast, and precise method to study protein–nucleic acid interactions in solution. Integrated quality controls help to obtain optimal data quality. MST consumes low sample material and allows to work at free choice of buffer conditions, making measurements in serum, whole blood, or cell lysate possible [12, 13].

4 Notes

1. The detection of low fluorophore concentrations requires high excitation light intensities (LED power > 75), which can cause significant photobleaching during the experiment and

introduce additional noise to the binding signal. Use the NanoTemper Anti-Photobleaching kit to prevent photobleaching throughout the experiment and to optimize the binding signal.

2. The concentration finder tool implemented in the NT. Control and NT. Analysis software simulates binding data and helps finding the right concentration range for the dilution series. It is either possible to simulate how the binding curve will look like at a certain K_d or at a certain "K_d Interval," if only a range of the binding affinity is known. Ideally, the concentration of ligand in the dilution series should be chosen in such a way that at least three data points are present in both the bound and unbound plateau of the binding curve.
3. For successful MST experiments, it is imperative that all molecules are free in solution. Some biomolecules however tend to stick to glass surfaces. Adsorption of fluorescent molecules to capillary walls can be identified by irregular capillary profiles in the initial capillary scan. Bumpy, flattened, or U-shaped capillary profiles indicate adsorption. *See* Fig. 3a for example profiles. To prevent these "sticking effects" several capillary types—each coated with different passivizing chemicals—are available. These capillaries should be tested for their suitability prior to binding experiments (*compare* Fig. 3a).
4. If the fluorescence intensity in the capillaries is below 200 counts, increase the concentration of labeled molecule or increase the LED power. If an increased laser power leads to significant photobleaching, please refer to **Note 1**.
5. Variations in fluorescence intensities in the capillary scan larger than 10% can be caused by (a) pipetting errors, (b) aggregation of labeled protein, (c) adsorption to capillary walls or labware, or (d) fluorescence quenching by the ligand.
   (a) To improve pipetting accuracy, make sure to use the exact same buffer of ligand stock and assay buffer. Mix the solution in each step at least eight times by pipetting up and down, do not introduce air bubbles while mixing, and do not use the "blow-out" function of your pipette.
   (b) Strategies to minimize aggregation are discussed in **Note 7**.
   (c) Adsorption of capillary walls can be identified and minimized as described in **Note 3**.
   (d) Fluorescence quenching by the ligand results in a systematic rather than a random fluorescence change. To test whether fluorescence loss at increasing ligand concentrations is caused by ligand binding or by ligand-induced denaturation/adsorption of the labeled protein, perform

the "SD-Test": Prepare two tubes each containing 10 μL of a 2× SD mix (4% SDS, 40 mM DTT). Carefully remove 10 μL of tubes 1 and 16 and transfer to the tubes containing the SD mix, mix well, and incubate for 5 min at 95 °C to denature the protein. Fill both samples into two standard capillaries each and measure the fluorescence intensity. In case of ligand-induced quenching, the fluorescence of denatured protein should be identical for both samples. If you observe a difference in fluorescence intensity for tubes 1 and 16, material was lost either by aggregation and subsequent centrifugation or by unspecific adsorption at the tube walls.

6. A total volume of 2 nL is heated by the infrared laser. The range of the temperature gradient depends on the MST power, and spans 2 °C (MST power 20%) to 6 °C (MST power 60%).
7. Aggregation of your protein can be prevented by adding detergents to the assay buffer (0.005–0.1% Tween 20, 0.01–0.1% Pluronic F127 or similar), by adding >0.5 mg/mL of stabilizing proteins such as BSA, and/or by centrifugation for >10 min at 22,000 × *g* prior to the experiment. Aggregates can also be identified by "bumpy" MST traces during the experiment (*see* Fig. 3b).
8. The NanoTemper analysis software offers two curve fit options: The fit function for K_d from the law of mass action and the fit function for EC50 from the Hill equation.

K_d from law of mass action:

$$A + T \Leftrightarrow AT.$$

$$\text{Fractionbound} = \frac{1}{2c_A}\left(c_T + c_A + K_d - \sqrt{(c_T + c_A + K_d)^2 - 4c_T c_A}\right),$$

K_d: dissociation constant, to be determined.

c_{AT}: concentration of formed complex.

c_A: constant concentration of molecule A (fluorescent), known.

c_T: concentration of titrated molecule T.

Please note that the fitting model from the law of mass action describes your data correctly when a molecule A interacts with a molecule B using one binding site or using multiple binding sites with the same affinity.

EC50 from the Hill equation:

$$A + nT \Leftrightarrow AT_n.$$

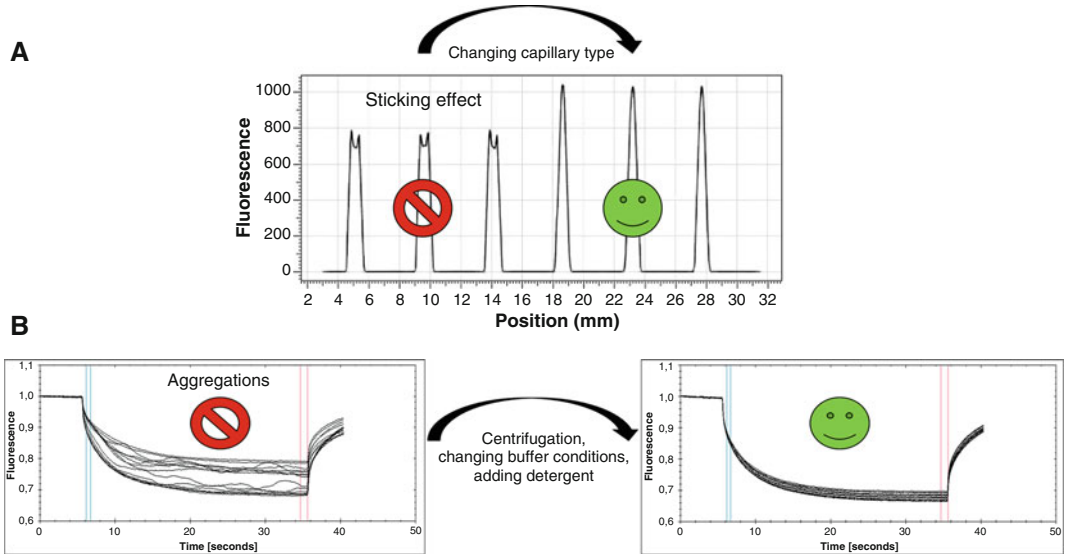


Fig. 3 Assay optimization for MST experiments. (**a**) Protein adsorption and sticking effects can be detected in the capillary scan. Irregular capillary shapes such as flattened or U-shaped peaks in the capillary scan indicate sticking of the sample material to the glass surface (*left*). In the example shown here, premium capillaries prevent molecule adsorption, resulting in a regular capillary shape (*right*). (**b**) The normalized fluorescence serves as quality control as it detects aggregates. "Bumpy" MST time traces indicate aggregation effects (*left*). Optimization of reaction buffers (e.g., including detergents such as Pluronic F127 or varying pH values or salt concentration) can improve the solubility of the molecules. Larger aggregates can be efficiently removed by centrifugation. Once aggregation is eliminated, the MST traces of identical samples should be indistinguishable as in the shown example (*right*)

$$\text{Fraction bound} = \frac{1}{1 + (\text{EC}_{50}/c_T)^n},$$

c_T: provided concentration of titrated molecule T.

The Hill model can be used to determine the EC50 value, which is the concentration of titrant where 50% of the fluorescent molecule is bound. Please keep in mind that the EC50 value is not a physical *constant* like the K_d, but an apparent measure of affinity for one particular experiment, which strongly depends on the used concentrations. For multivalent interactions, the Hill coefficient can provide information about the cooperativity of binding events.

References

1. Grummt I, Maier U, Ohrlein A, Hassouna N, Bachellerie JP (1985) Transcription of mouse rDNA terminates downstream of the 3′ end of 28S RNA and involves interaction of factors with repeated sequences in the 3′ spacer. Cell 43(3 Pt 2):801–810
2. Evers R, Grummt I (1995) Molecular coevolution of mammalian ribosomal gene terminator sequences and the transcription termination factor TTF-I. Proc Natl Acad Sci U S A 92 (13):5827–5831

3. Langst G, Becker PB, Grummt I (1998) TTF-I determines the chromatin architecture of the active rDNA promoter. EMBO J 17 (11):3135–3145. doi:10.1093/emboj/17.11.3135
4. Diermeier SD, Nemeth A, Rehli M, Grummt I, Langst G (2013) Chromatin-specific regulation of mammalian rDNA transcription by clustered TTF-I binding sites. PLoS Genet 9 (9):e1003786. doi:10.1371/journal.pgen.1003786
5. Duhr S, Braun D (2006) Why molecules move along a temperature gradient. Proc Natl Acad Sci U S A 103(52):19678–19682. doi:10.1073/pnas.0603873103
6. Duhr S, Arduini S, Braun D (2004) Thermophoresis of DNA determined by microfluidic fluorescence. Eur Phys J E Soft Matter 15 (3):277–286. doi:10.1140/epje/i2004-10073-5
7. Baaske P, Wienken CJ, Reineck P, Duhr S, Braun D (2010) Optical thermophoresis for quantifying the buffer dependence of aptamer binding. Angew Chem Int Ed Eng 49 (12):2238–2241
8. Ludwig C (1856) Diffusion zwischen ungleich erwärmten Orten gleich zusammengesetzter Lösungen. Sitzungber Bayer Akad Wiss Wien Math, Naturwiss Kl 20
9. Seidel SA, Wienken CJ, Geissler S, Jerabek-Willemsen M, Duhr S, Reiter A, Trauner D, Braun D, Baaske P (2012) Label-free microscale thermophoresis discriminates sites and affinity of protein-ligand binding. Angew Chem Int Ed Eng 51(42):10656–10659
10. Schubert T, Pusch MC, Diermeier S, Benes V, Kremmer E, Imhof A, Langst G (2012) Df31 protein and snoRNAs maintain accessible higher-order structures of chromatin. Mol Cell 48(3):434–444. doi:10.1016/j.molcel.2012.08.021
11. Zillner K, Jerabek-Willemsen M, Duhr S, Braun D, Langst G, Baaske P (2012) Microscale thermophoresis as a sensitive method to quantify protein: nucleic acid interactions in solution. Methods Mol Biol 815:241–252
12. Wienken CJ, Baaske P, Rothbauer U, Braun D, Duhr S (2010) Protein-binding assays in biological liquids using microscale thermophoresis. Nat Commun 1:100
13. Seidel SA, Dijkman PM, Lea WA, van den Bogaart G, Jerabek-Willemsen M, Lazic A, Joseph JS, Srinivasan P, Baaske P, Simeonov A, Katritch I, Melo FA, Ladbury JE, Schreiber G, Watts A, Braun D, Duhr S (2013) Microscale thermophoresis quantifies biomolecular interactions under previously challenging conditions. Methods 59(3):301–315

Chapter 11

Establishment of the CRISPR/Cas9 System for Targeted Gene Disruption and Gene Tagging

Eric Ehrke-Schulz, Maren Schiwon, Claudia Hagedorn, and Anja Ehrhardt

Abstract

CRISPR/Cas9 RNA-guided nucleases refashioned in vivo gene editing approaches for specific gene disruption, gene correction, or gene addition. Moreover, chimeric Cas9 proteins can be applied to direct fused *cis*-acting effector protein domains, enzymes, or fluorescent markers to DNA to target sequences to regulate gene expression, to introduce epigenetic changes, or to fluorescently label DNA sequences of interest. Here we show how to design guide RNAs for specific DNA targeting. We provide a protocol to customize the CRISPR/Cas9 machinery encoded on commercially available plasmids and present how to test the targeting efficiency of Cas9 with a target-specific gRNA by testing mutation induction efficiency. To exemplify related applications we provide a guideline of how to apply the CRISPR/Cas9 technology for gene labeling.

Key words CRISPR/Cas9, gRNA design, Gene disruption, T7E1 assay, Gene tagging

1 Introduction

The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/Cas9 system revolutionized the field of designer nuclease-based gene editing. The CRISPR/Cas9 system is widely used for various genome editing approaches in cultured cells and living organisms and was broadly explored for preclinical applications. This two component system is composed of the RNA-guided Cas9 endonuclease that acts in cooperation with a chimeric guide RNA (gRNA) mediating the sequence-specific binding to its complementary target sequence preceding a protospacer adjacent motif (PAM) within the genome [1]. Upon binding to its target sequence the Cas9 enzyme introduces DNA double strand breaks (DSB) three base pairs upstream of the PAM motif. In absence of a DNA that is complementary to the interrupted locus, eukaryotic cells repair DNA DSB via the error prone nonhomologous end joining (NHEJ) pathway. This leaves small insertions or deletions or a combination of both at the repaired DSB site. Thus NHEJ of

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_11,

DNA DSB often results in disruption of the open reading frame at the target site. Several different versions of Cas9 containing mutations rendering the protein a nickase or having no nuclease function have been developed broadening the potential applications of this system.

Cas9 without nuclease function can be fused with other *cis*-acting proteins such as transcription factors, methylases, or fluorescent proteins [2, 3]. Depending on the kind of fused protein CRISPR/Cas9 can be used for locus-specific regulation of gene expression, epigenetic modification, or tagging of DNA loci for imaging purposes. Using the CRISPR/Cas9 system for visualization of endogenous genomic loci opens new perspectives in studying genome function regulation, spatial-temporal genome organization, and interaction with subnuclear structures. Since visualization of genomic loci solely requires binding of the Cas9, a "dead" Cas9 (dCas9) with inactivated RuvC1 (amino acid substitution D10A) and HNH (amino acid substitution H840A) [4] is fused to either fluorescent proteins such as green fluorescent protein (GFP) or red fluorescent protein (RFP) [5], or to a Halo tag that can covalently and efficiently be labeled with Halo ligands conjugated to a variety of organic fluorescent dyes [6, 7]. Applying ortholog Cas9 variants from different bacterial species, e.g., *Streptococcus thermophilius* (St1) and *Neisseria meningitides* (Nm), enables for multicolor labeling [5]. However, in contrast to conventional DNA FISH, usage of a fused dCas9 (e.g., dCas9-GFP) does not require antibody staining, thus reducing the risk of unspecific binding on the one hand, but, on the other hand, lacks any signal amplification. Therefore, genomic regions with >1000 repeats (e.g., telomeres, pericentromeric regions, major satellites) and genetic regions containing repetitive sequences such as the human MUC4 intron3 (approx. 90 repeats), can efficiently be labeled with only one target gRNA [5, 8]. Arbitrary non-repetitive gene loci in contrast may require 36–73 different gRNA for efficient labeling [8].

Due to its simple gRNA design and easy cloning procedure for customization for desired applications, the CRISPR/Cas9 system is easier to handle than transcription activator-like effector nucleases (TALENs) and artificial zinc finger nucleases (ZFN) [9]. Here we provide a simple guideline how to design gRNAs for the desired application. We show how to customize the CRISPR/Cas9 machinery encoded on commercially available plasmids and how to test the targeting efficiency of a respective gRNA by testing their efficiency to introduce mutations at the desired DNA target sequence.

To design specific gRNAs to guide binding of the CRISPR/Cas9 machinery to the target sequence of choice for both, labeling and genome editing, online prediction tools help to select possible gRNA binding sites and to highlight target specificity using a

scoring system. Furthermore, they discover possible off-targets within the respective genome and predict likeliness of the respective gRNAs to mediate Cas9 binding at these potential off-target sites. Predicted gRNA oligonucleotides can be ordered via commercial primer synthesis services. Single-stranded gRNA oligonucleotides are rendered double stranded and inserted into respective cloning sites of available CRISPR/Cas9 plasmids by applying a simple cut and ligation protocol [10]. Customized CRISPR/Cas9 plasmids are transfected into the cultured cells of choice and targeting efficiency is evaluated by mutation detection using the T7E1 assay. This assay is based on heteroduplex formation between mutated and non-mutated PCR products of the target locus from CRSIPR/Cas9-transfected cells and subsequent cleavage of heteroduplexes by the mismatch-sensitive T7 endonuclease I.

2 Materials

2.1 CRISPR/Cas9 Vector Construction

1. Computer with internet access and an e-mail address.
2. Single-stranded gRNA forward- and reverse gRNA oligonucleotides 100 μM (*see* **Note 1**), 10× T4 Ligation Buffer (New England Biolabs, NEB) (*see* **Note 2**), T4 polynucleotide kinase (PNK, NEB), Thermocycler (*see* **Note 3**), ddH_2O.
3. pX330-U6-Chimeric_BB-CBh-hSpCas9 plasmid (Addgene) [10] (*see* **Note 4**), 10× Cutsmart Buffer (NEB), Dithiothreitol (DTT, 10 mM), ATP (10 mM), *Bbs*I restriction enzyme (NEB), T4 DNA ligase (NEB).
4. Fusion-dCas9 expression plasmids (e.g., pHAGE-TO-nmdCas9-3XmCherry, pHAGE-TO-nls-st1dCas9-3nls-3XGFP-2nls; Addgene) and gRNA expression plasmids (pLH-nmsgRNA1.1, pLH-stsgRNA2.1; Addgene) [5].
5. 10× Buffer 4 (NEB), ATP (10 mM), exonuclease V (NEB), water bath or incubator, EDTA solution (1 M).
6. Commercially available column cleanup kit for DNA cleanup from reaction, alternatively potassium acetate (3 M, pH 8), Ethanol 100%, Ethanol 70%, ddH_2O.
7. Competent *E. coli*, LB plates containing ampicillin (50 μg/mL), bacterial incubator, LB Medium containing ampicillin (50 μg/mL), shaking incubator or thermomixer.
8. Restriction enzyme of choice and appropriate buffer.

2.2 Testing of CRISPR/Cas9 Constructs

1. A cell line of choice, appropriate cell culture medium (*see* **Notes 5** and **6**), 24-well tissue culture plate or chamber slides for CRISPR/Cas9 imaging, CRISPR/Cas9 expression plasmid (>200 ng/μL), target plasmid (>200 ng/μL) (optional) (*see*

Note 6), $CaCl_2$ (2.5 M), 2× HEPES-buffered saline (HBS) (140 mM NaCl, 5 mM $Na_2HPO_4 \cdot 2H_2O$, 50 mM HEPES).

2. Table top centrifuge, cell lysis buffer (10 mM Tris–Cl pH 8.0, 100 mM EDTA pH 8.0, 50 mM NaCl, 0.5% SDS, 20 μg/mL RNAse, 100 μg/mL Proteinase K) (*see* **Note 7**), thermomixer.
3. Phenol/chloroform/isoamyl alcohol (25:24:1), ice-cold ethanol (100%), sodium acetate (3 M pH 5.2), ethanol (70%), 1× TE buffer (10 mM Tris–HCl, 1 mM EDTA, pH 7.6).
4. PCR reagents, primers (*see* **Note 8**), Thermocycler.
5. Agarose gel (2%), 1× TAE buffer (40 mM TRIS, 1 mM EDTA-Na_2-salt, 40 mM acetic acid), gel loading dye, 1000 bp molecular weight marker, electrophoresis device, gel documentation system or UV table.
6. 10× Buffer 2 (NEB), T7 Endonuclease I (T7E1, NEB), gel loading dye containing SDS and EDTA (e.g., 6× purple gel loading dye (NEB), Ice.
7. Agarose gel (2%), 1× TAE buffer, 1000 bp molecular weight marker, agarose gel electrophoresis device, gel documentation system or UV table.
8. 1× Phosphate buffered saline (PBS) (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4), 4% paraformaldehyde/1 × PBS, 0.5% NP-40/1 × PBS, 4′,6-diamidino-2-phenylindole (DAPI).
9. Fluorescence microscope for imaging.

3 Methods

3.1 gRNA Binding Site Prediction and gRNA Oligonucleotide Design

1. Open http://crispr.mit.edu/ (*see* **Note 9**), enter your E-mail address and name your search.
2. Specify the target genome (originating organism).
3. Enter the DNA sequence of your intended target locus and submit your query.
4. Close the browser or tab (*see* **Note 10**), wait for an E-mail, then click on the link to your results. Your result will open.
5. Choose 3–5 gRNA binding sites with high binding specificity and minimal potential off-target sites to proceed with further gRNA oligonucleotide design for insertion into the CRISPR/Cas9 expression plasmid (*see* **Note 11**).
6. Add the sequence CACC to the 5′ end of the forward gRNA oligo (Fig. 1) (*see* **Note 12**).
7. Add AAAC to the complementary sequence of the forward binding site, which is then the reverse gRNA oligo (Fig. 1).

Fig. 1 gRNA oligonucleotide design. (**a**) Example of a potential target locus sequence with a predicted gRNA binding site highlighted by the *light blue box*. The PAM motif (NGG) is marked in *darker blue*. Oligonucleotides are marked as *pink arrows*. (**b**) Oligonucleotides that have to be synthesized. (**c**) Final annealed double-stranded oligonucleotide for insertion into the *Bbs*I oligonucleotide insertion site within the gRNA expression unit of the CRISPR/Cas9 expression plasmid

Order your oligonucleotides at your primer synthesis service of choice.

8. For imaging add the sequence ACCG to the 5′ end of the forward gRNA oligo; add CACC (pLH-nmsgRNA1.1) or AGAC (pLH-stsgRNA2.1) to the complementary sequence of the forward binding site.

3.2 Phosphorylation and Annealing of gRNA Oligonucleotides

1. Mix 1 μL of forward gRNA oligonucleotide (100 μM) and 1 μL of the corresponding reverse gRNA oligonucleotide (100 μM) with 1 μL 10× T4 Ligation Buffer, 0.5 μL T4 PNK and add ddH_2O to a final volume of 10 μL.
2. Incubate at 37 °C for 30 min, followed by incubation at 95 °C for 5 min and finally cool down to 25 °C at 0.1 °C/s using a thermocycler.

3.3 Insertion of Phosphorylated and Annealed gRNA Oligonucleotide Into the CRISPR/Cas9 Expression Vector

1. Dilute the annealed oligonucleotide 250-fold using ddH_2O.
2. Set up the digestion-ligation reaction containing pX330 (Fig. 2) or other backbone vector (100 ng), 2 μL phosphorylated and annealed oligonucleotide duplex (1:250 dilution), 2 μL 10× Cutsmart Buffer (NEB), 2 μL DTT, 1 μL ATP, 1 μL *Bbs*I (NEB), 0.5 μL T4 DNA ligase (NEB) and ddH_2O to a final volume of 20 μL.
3. Incubate the ligation reaction in a thermocycler: 37 °C for 5 min followed by 23 °C for 5 min. Cycle these steps six times followed by incubation at 4 °C until further processing.

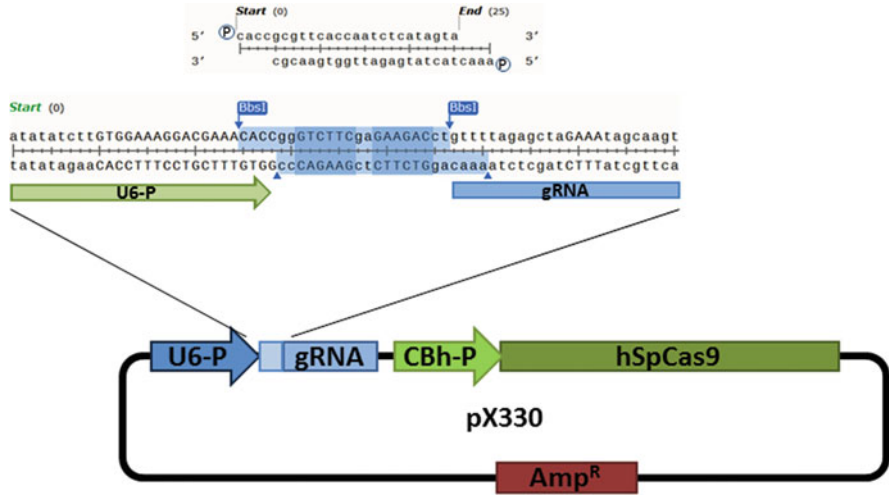


Fig. 2 Insertion of phosphorylated double-stranded gRNA oligonucleotide into the *Bbs*I insertion site of the gRNA expression unit of the CRISPR/Cas9 expression vector. The pX330 plasmid with the hSpCas9 and gRNA expression cassettes is schematically shown. Note that the gRNA insertion site is enlarged to illustrate customization of the gRNA. The sequence that needs to be replaced to customize the gRNA specificity is shown in *light blue*. *Bbs*I recognition sites are indicated by *blue boxes*. *Bbs*I cutting sites are indicated by *blue arrowheads*

3.4 Exonuclease Treatment to Prevent Unwanted Recombination Products

1. Combine 11 μL ligation reaction with 1.5 μL Buffer4 (NEB) 1.5 μL ATP (10 mM), 1 μL exonuclease V (NEB). Incubate at 37 °C for 30 min. Then add EDTA to a final concentration of 11 mM to stop the reaction.
2. Heat inactivate at 70 °C for 30 min (optional). Then clean-up treated samples by column purification using a commercial reaction clean-up kit or ethanol precipitation as described in **steps 4–6** of subheading 3.7.

3.5 Transformation Into E. coli and Clone Verification

1. Transform 1–2 μL of the final product into competent *E. coli* and plate bacteria on LB plates containing ampicillin (50 μg/mL). Incubate at 37° for 16–24 h.
2. Pick colony and inoculate overnight culture. On the next day purify plasmids using a commercially available plasmid DNA isolation kit or any other method of choice.
3. Verify clones using an appropriate restriction enzyme digest (*see* **Note 13**). Sequence your plasmids.

3.6 Transfection of CRISPR/Cas9 Constructs

Transfect plasmids containing gRNA and Cas9 expression cassettes and optional the target sequence into an appropriate cell line (*see* **Note 14**). Use calcium phosphate-mediated transfection [11] or any other appropriate transfection method that results in high transfection efficiencies (*see* **Note 15**).

1. Twenty-four hours before transfection seed ~70,000 cells in, e.g., a 24-well plate to reach 30–60% (*see* **Note 16**) confluency the next day. For CRSIPR/Cas imaging grow cells on coated cover slips (22 × 22 mm) or chamber slides of appropriate size.
2. Three hours prior to transfection exchange cell culture media with fresh media.
3. Prepare transfection mixture combining 20 μL of $CaCl_2$ with a total amount of 1 μg of plasmid DNA and fill up to 25 μL with ddH_2O. When using two different expression plasmids for fused-dCas9 and gRNA constructs, mix plasmids at a ratio of 200 ng for the dCas9 encoding plasmid with 800 ng of gRNA plasmid DNA.
4. Add to 25 μL of the $CaCl_2$-DNA solution to 25 μL 2× HBS. Mix thoroughly and incubate for 30 min. Then mix again and immediately add the 50 μL $CaCl_2$-DNA-HEPES solution dropwise onto the cells of one well of a 24-well plate and swirl the plate. Change medium next day.
5. Incubate for 48–72 h and then harvest cells for genomic DNA isolation as described under Subheading 3.7 or imaging procedures.

3.7 Genomic DNA Isolation

1. Harvest cells by flushing of the dish with the cell culture media and pellet the cells by centrifugation at 300 × *g* for 3 min (*see* **Note 17**).
2. Discard supernatant and lyse cell pellet by adding 40 μL lysis buffer and incubate at 50 °C for 3 h and shake at ~1100 rpm.
3. Add 400 μL ddH_2O and an equal volume Phenol/Chloroform/Isoamyl alcohol (25:24:1) and centrifuge at 14,000 × *g* for 2 min.
4. Carefully transfer the upper phase to a fresh tube, add an equal volume of Phenol/Chloroform/Isoamyl alcohol (25:24:1), and centrifuge at 14,000 × *g* for 2 min. Repeat **step 4** until there is no protein left in the interphase.
5. Transfer upper phase to a fresh tube and precipitate DNA by adding 0.1 volumes sodium acetate and 2.5 volumes ice-cold ethanol (100%). Centrifuge at 20,000 × *g* for 10 min.
6. Discard the supernatant without disturbing the DNA pellet. Add 500 μL ethanol (70%). Centrifuge again at 20,000 × *g* for 5 min.
7. Discard supernatant without disturbing the DNA pellet. Air-dry the DNA pellet and resolve DNA in 100 μL 1× TE buffer or ddH_2O.

3.8 Mutation Detection with T7 Endonuclease I

1. To amplify the genomic locus surrounding your gRNA target site set up a PCR reaction in a total volume of 50 µL using 3 µL of genomic DNA isolated from CRISPR/Cas9 plasmid-transfected cells. Apply at least 35 amplification cycles to ensure efficient amplification of the target locus (*see* **Note 18**).
2. Precipitate the PCR product as described in **steps 3–7** of Subheading 3.7. Dissolve the pelleted PCR product in 20 µL ddH_2O and analyze 5 µL on an agarose gel to evaluate amplification and DNA quality.
3. If PCR products appear clean in the agarose gel, set up heteroduplex formation reaction by mixing 9 µL of purified PCR product (*see* **Note 19**) with 1 µL Buffer 2 (NEB). Induce heteroduplex formation in a thermocycler using following conditions: incubate at 95 °C for 2 min, then cool to 85 °C using a cooling rate of 2 °C/s, then cool to 25 °C using a cooling rate of 0.1 °C/s and finally incubate sample at 16 °C until further processing.
4. Add 5 µL T7E1 master mix (4 µL H_2O, 0.5 µL Buffer 2 (NEB) and 0.5 µL T7E1 enzyme (NEB). Incubate at 37 °C for 15 min. Put on ice and stop reaction by adding 3 µL of loading dye.
5. Separate digest on an agarose gel (2%) (*see* **Note 20**) and analyze bands using a gel imaging instrument. In the case of successfully induced gene modification, cleavage products with lower molecular weight than the original PCR product can be detected (Fig. 3). Estimate target disruption efficiency using the formula published by Miller et al. [12].

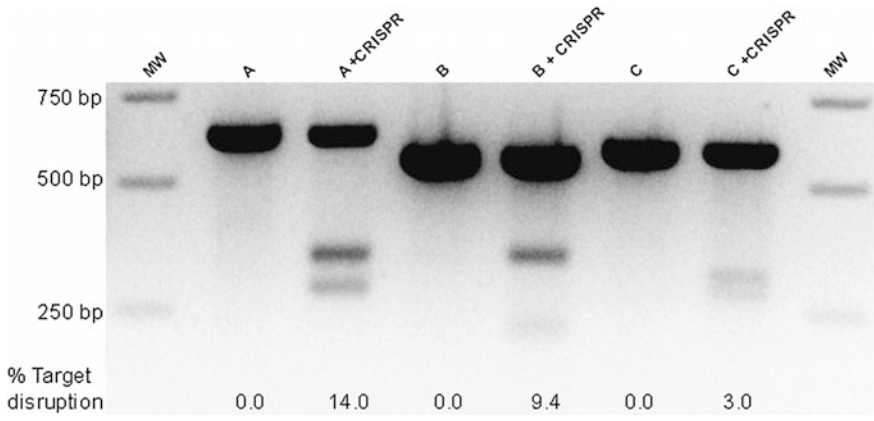


Fig. 3 Evaluation of T7E1 assay by agarose gel electrophoresis. Three examples of CRISPR/Cas9-mediated mutation induction (*A*, *B*, *C*) are displayed. Lanes *A*, *B*, and *C* show PCR products of untreated cells, and lanes *A* + *CIRISPR*, *B* + *CRISPR*, and *C* + *CRISPR* show respective cleavage products of mutation insertion. The measured target disruption frequency is indicated below (% target disruption). *MW* molecular weight marker

3.9 Imaging Procedure

1. For imaging fix cells in 4% paraformaldehyde/1 × PBS for 10 min. When required, permeabilize cells with 0.5% NP-40/1 × PBS for 10 min and counterstain nuclei with DAPI for 2–5 min. Subsequently wash with twice 1 × PBS for 5 min.
2. Apply mounting medium, cover with a coverslip and seal with colorless nail polish. For live cell imaging, cells grown on chamber slides can be imaged using a microscope stage incubation chamber maintained at 37 °C.
3. Use an appropriate fluorescence microscope to visualize cells.

4 Notes

1. Lyophilized single-stranded gRNA forward and reverse gRNA oligonucleotides can be ordered at any commercial supplier and diluted in ddH_2O to the final concentration.
2. In the protocols provided in this chapter we use enzymes purchased from New England Biolabs (NEB) with which we established all methods. Alternatively, enzymes can be ordered at any commercial supplier, but be aware to use the appropriate buffer supplied with the respective enzymes to ensure its fidelity. We do not recommend mixing enzymes from two different suppliers within the same reaction.
3. The thermocycler should allow for slow cooling at 0.1 °C/s or slower.
4. This plasmid contains two expression cassettes, humanized *S. pyogenes* hSpCas9 and the chimeric gRNA expression unit (Fig. 2). The vector can be digested using *Bbs*I, and a pair of annealed oligonucleotides can be cloned into the gRNA expression unit. The oligonucleotides are designed based on the predicted target site sequence (20 bp). This version of the plasmid contains a longer fragment of the tracrRNA (+85 nt). For the protocol described here any other CRISPR/Cas9 vectors containing a *Bbs*I cloning site for oligonucleotide insertion within the gRNA expression unit can be used.
5. We recommend using the cell line that is relevant for your desired application.
6. If the target site of a specific gRNA is not present in your cultured cells, such as episomal DNA like viral genomes or plasmid replicons, the user can co-transfect plasmid DNA containing the desired target locus.
7. We recommend adding the enzymes prior to buffer application as they may lose activity when stored within the buffer for long time.

8. The user is free to utilize a polymerase and PCR reagents of choice. We recommend designing primers to surround the intended gRNA target sites resulting in a PCR product of approximately 500 bp or larger in size with the gRNA target site located in the middle of the PCR product.
9. Here we use the "CRISPR DESIGN" tool for the gRNA prediction. There are numerous other online tools. The user is free to choose.
10. An E-mail will be send to you when the search is completed. You access the output of your search via a link that is send by E-mail. You can access your search results for 1 month.
11. To choose gRNA sequences with high specificity, choose a sequence with a high score. To choose a sequence with potentially low number of target sites refer to the predicted number of off-target sites for the respective sequence. Here the user has to find the optimal balance between specificity and potential off-target sites.
12. The gRNA sequence for further gRNA oligonucleotide design only contains the gRNA 20 bp sequence preceding the PAM motif. The PAM motif is not included within the oligonucleotide. PAM choices are NGG (Sp Cas9), NNNNGATT (Nm Cas9), and NNAGAAW (St Cas9).
13. The *Bbs*I restriction enzyme recognition site of the original vector should be absent in final constructs. To analyze candidate clones by restriction enzyme digest perform a double digest using *Bbs*I and a second restriction enzyme that cuts only once within your plasmid approximately 1 kb downstream or upstream from the gRNA oligonucleotide insertion. Plasmids of correct clones should be linearized and be visible on an agarose as a single band. Unaltered vectors show an additional band of approximately 1 kb.
14. High quality plasmid DNA is recommended. A good indicator of DNA purity is the ratio of absorbance at 260 nm (A260) to 280 nm (A280). A DNA solution with an A260/A280 ratio of 1.8 or greater is desirable.
15. Transfection efficiencies can be increased in many cell types by additional treatments after the primary exposure of cells to calcium phosphate-precipitated DNA. The most effective and routinely used agents are glycerol, dimethyl sulfoxide (DMSO), chloroquine, and sodium butyrate [12].
16. An optimal plating density produces a nearly confluent dish when the cells are harvested after 48 h. Here longer incubation times favor mutation induction, so cells are usually harvested at a later time point. Nevertheless, cells should not be seeded too thin, so that the viability is not negatively influenced.

17. Here we use the alkaline lysis method for genomic DNA isolation; the user may also use any other protocol or commercially available kits suitable to isolate genomic DNA from cultured cells.
18. The user is free to choose an appropriate polymerase/PCR system. Primers used for this PCR should allow amplifying the genomic locus surrounding the gRNA binding site resulting in a PCR product of 500–1000 bp in which the gRNA target sequence is located roughly in the middle of the PCR product.
19. Use sufficient amounts of PCR product for the T7E1 assay. Usually ½ of the PCR reaction is sufficient.
20. As a rule, running the gel for at least 1 h at medium voltage leads to well-separated bands. However, conditions can be optimized.

Acknowledgments

This work was supported by the Witten/Herdecke University internal research promotion Grant No. IFF2017-12 and the German Duchenne Foundation "Aktion Benni & Co." starting grant to E.E.-S.

References

1. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337 (6096):816–821. doi:10.1126/science.1225829
2. Dambournet D, Hong SH, Grassart A, Drubin DG (2014) Tagging endogenous loci for live-cell fluorescence imaging and molecule counting using ZFNs, TALENs, and Cas9. Methods Enzymol 546:139–160. doi:10.1016/B978-0-12-801185-0.00007-6
3. Tanenbaum ME, Gilbert LA, Qi LS, Weissman JS, Vale RD (2014) A protein-tagging system for signal amplification in gene expression and fluorescence imaging. Cell 159(3):635–646. doi:10.1016/j.cell.2014.09.039
4. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA (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
5. Ma H, Naseri A, Reyes-Gutierrez P, Wolfe SA, Zhang S, Pederson T (2015) Multicolor CRISPR labeling of chromosomal loci in human cells. Proc Natl Acad Sci U S A 112 (10):3002–3007. doi:10.1073/pnas.1420024112
6. Deng W, Shi X, Tjian R, Lionnet T, Singer RH (2015) CASFISH: CRISPR/Cas9-mediated in situ labeling of genomic loci in fixed cells. Proc Natl Acad Sci U S A 112(38):11870–11875. doi:10.1073/pnas.1515692112
7. Encell LP, Friedman Ohana R, Zimmerman K, Otto P, Vidugiris G, Wood MG, Los GV, McDougall MG, Zimprich C, Karassina N, Learish RD, Hurst R, Hartnett J, Wheeler S, Stecha P, English J, Zhao K, Mendez J, Benink HA, Murphy N, Daniels DL, Slater MR, Urh M, Darzins A, Klaubert DH, Bulleit RF, Wood KV (2012) Development of a dehalogenase-based protein fusion tag capable of rapid, selective and covalent attachment to customizable ligands. Curr Chem Genomics 6:55–71. doi:10.2174/1875397301206010055
8. Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, Li GW, Park J, Blackburn EH, Weissman JS, Qi LS, Huang B (2013) Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell 155(7):1479–1491. doi:10.1016/j.cell.2013.12.001

9. Maeder ML, Gersbach CA (2016) Genome-editing Technologies for Gene and Cell Therapy. Mol Ther 24(3):430–446. doi:10.1038/mt.2016.10
10. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339(6121):819–823. doi:10.1126/science.1231143
11. Sambrook J, Russell DW (2001) Molecular cloning: a laboratory manual, vol Bd. 1–3. Cold Spring Harbor Laboratory Press, New York
12. Miller JC, Holmes MC, Wang J, Guschin DY, Lee YL, Rupniewski I, Beausejour CM, Waite AJ, Wang NS, Kim KA, Gregory PD, Pabo CO, Rebar EJ (2007) An improved zinc-finger nuclease architecture for highly specific genome editing. Nat Biotechnol 25(7):778–785. doi:10.1038/nbt1319

Part III

RNA Analysis

Chapter 12

Holistic and Affordable Analyses of MicroRNA Expression Profiles Using Tagged cDNA Libraries and a Multiplex Sequencing Strategy

Patrick P. Weil, Yan Jaszczyszyn, Anne Baroin-Tourancheau, Jan Postberg, and Laurence Amar

Abstract

Small and long noncoding RNAs (ncRNAs) are key regulators of gene expression. Variations in ncRNA expression patterns can consequently affect the control of many cellular processes. Not just since 2006, when Andrew Z Fire and Craig C Mello were jointly awarded The Nobel Prize in Physiology or Medicine for their discovery of RNA interference, great efforts were undertaken to unleash the biomedical applicability of small noncoding RNAs, in particular microRNAs. With the technological evolution of massive parallel sequencing technologies over the last years, which now are available for an increasing number of scientists, there is a demand for comprehensible and efficient workflows reliable even for unique and valuable clinical specimens. Here we describe a highly reproducible low-cost protocol for analyses of miRNA expression patterns using tagged cDNA libraries and a multiplex sequencing strategy following an Illumina-like protocol. This protocol easily allows the identification of expression differences from samples of tissues of 1–2 mm^3 and fluids of 50–200 μL. We further provide entry points into useful computational biology applications, whose target groups explicitly involve non-bioinformaticians.

Key words microRNA, miRNome, Multiplex sequencing

1 Introduction

Several years after the discovery of the RNA interference (RNAi) process [1] it is well known that small noncoding RNAs (ncRNAs) such as microRNAs (miRNAs) as well as some long noncoding RNAs (lncRNAs) are key regulators of gene expression, which not only contribute to cellular differentiation and multiform developmental programs in the course of ontogeny, but might be also involved in pathophysiological processes linked with many complex diseases. Numerous studies have claimed the usability of miRNAs as biomarkers for personalized diagnostics, the deeper understanding of disease-related pathways and as potential therapeutic agents [2–4]. But up to date most practical uses still remain in

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_12,

their infancy. Currently and in future, this field's community will endeavor to uncover the biological relevance of miRNAs as key molecules in regulatory circuits and to identify deviations associated with diseases from a holistic, "miRNome"-wide point of view. The greatest challenge, however, will be the integration of these data with study results from all complementary levels of gene expression regulation. Those include genomics, epigenomics, and eventually proteomics—since it is the crosstalk between these regulatory entities that finally shapes cellular and organismal phenotypes between different species as well as between individual beings [5].

MiRNA expression profiles can be cell-type-, tissue-, or developmental stage-specific [6–9]. The biogenesis of miRNAs is a complex multistep process. Almost half of all miRNAs are encoded by genes (miR genes), which are transcribed by RNA polymerase II [10]. Others reside within introns. Their transcription can depend either on the host gene expression, or it is under the control of own specific promoters. Primary miRNA transcripts (pri-miRNA) display imperfect hairpin structures, which are recognized by the microprocessor complex consisting of the RNase III-enzyme Drosha and cofactor DGR8 [11, 12]. The microprocessor complex cleaves pri-miRNAs into hairpins of about 70 nucleotides in length (pre-miRNAs). These become exported from the nucleus into the cytoplasm and get cleaved under participation of another RNase III-enzyme, Dicer, and TRBP, eventually leading to the generation of imperfect duplexes of 22 base pairs [13, 14]. The slicing activity, which is inherent to many eukaryotic Argonaute proteins, leads to the depletion of the "passenger" strand from the duplex form of the microRNAs, whereas the remaining "guide"-strand remains bound to Argonaute (Ago) [15]. Single strands of miRNAs, once associated with Argonaute (Ago) constitute the catalytic core component of the active RNA-induced silencing complex (RISC). They thus participate in the posttranscriptional regulation of gene expression. The incorporated "guide" miRNA then directs RISC to its messenger RNA (mRNA) target(s). Animal miRNAs can recognize their targets via base pairing of only as few as 6–8 nucleotides in their 5′-seed region [16, 17]. Through sequence complementary, a single miRNA is predicted to target hundreds of different mRNA species [18, 19], but the question whether those putative interactions exist truly in vivo must be addressed experimentally. Eventually, functional interactions between miRNAs and the target sites within mRNAs—often in their 3′-untranslated regions (UTR)—impair mRNA translation (blocking) and/or stability (cleavage through slicing Argonautes) [15].

While our understanding about the mechanisms of miRNA biogenesis has already reached a relatively detailed level, the knowledge about miRNome-wide dynamics in health and disease, true and functional miRNA-mRNA target interactions, and the

interplay of the multiple gene regulatory circuits is still fragmentary. Here we provide a well-proven methodological pipeline for holistic miRNome analyses to an interdisciplinary community of researchers, who hopefully will contribute to expand the availability of relevant data in future.

1.1 MicroRNA Expression Profiling Using Illumina Sequencing Technology

Microarray and RT-qPCR technologies suffer from specific limitations: Only probe-related outputs can be monitored, cross-reactions produce false positive data and isomiRs (i.e., molecules differing by a few bases at their 3′ and 5′ ends) are not individually analyzed. Massive parallel sequencing technologies, in particular the Illumina technology, skip these limitations with currently several hundreds of millions of reads being analyzed per flowcell lane, making it the technology of choice for exhaustive investigation of miRNA expression. The primary goal of quantitatively investigating gene expression is to measure differences in the levels of transcripts between different samples. Note that read counts do not reflect the miRNA abundance in the original sample [20, 21]. The method described here is based on the Illumina® TruSeq™ Small RNA Sample preparation protocol which utilizes the advantage of the natural structure of mature miRNAs in animals, exhibiting a 5′-phosphate and a 3′-hydroxyl group. The method that we describe step by step allows home-construction of cDNA libraries at an affordable price with less than 50 euros per sample.

1.2 Multiplex Sequencing Strategy

Starting from frozen or fresh tissue, cell culture, FFPE or blood samples, total RNA is isolated by using a mixture of acidified phenol/guanidine thiocyanate and chloroform or column-based extraction methods. When having fresh tissue initial enzymatic digestion and cell separating steps using, i.e., magnetic activated cell sorting (such as Miltenyi Biotec's MACS) should be considered to obtain purified cell types. To assess the quantity of total RNA a Photometer (such as Implen's NanoPhotometer) can be used to measure the absorbances at 260–280 nm. A ratio of absorbance at 260–280 nm of ~2.0 is indicative of successful RNA purification. If the ratio is appreciably lower, it may indicate the presence of protein, phenol, or other contaminants that absorb strongly around 280 nm. In that case RNA concentration could be overestimated. It could be applicable to use a fluorimeter (such as Life Technologies' Qubit or Promega's Quantus) and a Bioanalyzer (such as Agilent's 2100 Bioanalyzer), because these methods are less sensitive to contaminants such as phenol or genomic DNA. On a fluorimeter the RNA concentration is measured by the absorbance of a fluorescent dye which is bound to the RNA. The Bioanalyzer is a microscale electrophoresis as well based on the absorbance of a fluorescent dye bound to the RNA and the different elution times of different sized RNA fragments. The so-called electropherogram of eukaryotic RNAs mainly shows two well-defined peaks

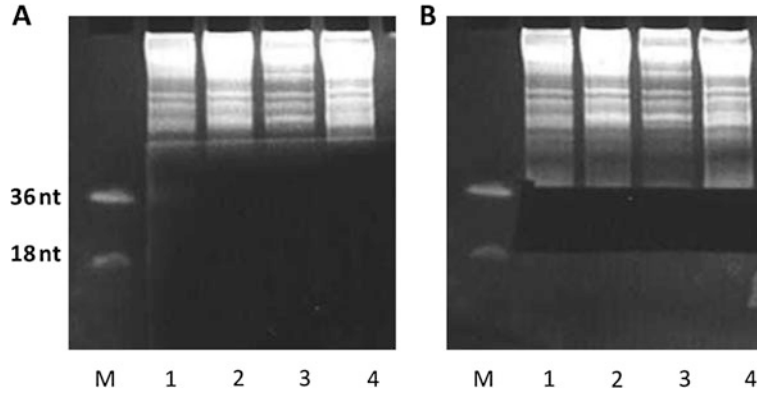


Fig. 1 Enrichment of small RNAs is essential in order to obtain high numbers of reads. (**a**) Selection of small RNAs using polyacrylamide gel electrophoresis. Only the ladder and the first approx. 5 cm of each sample (exhibiting enrichment of larger RNAs, such as t-RNAs, r-RNAs, and mRNA fragments) exposed to UV. (**b**) Areas containing RNA of 18–36 nucleotides cut from the gel separately for each sample

corresponding to the 18S ribosomal (rRNA) of 1.9 kb and 28S rRNA of 4.5 kb with a ratio of approximately 2:1. This total RNA ratio is calculated by taking the ratio of the area under the 18S and 28S rRNA peaks to the total area under the graph. The Bioanlayzer software uses algorithms based on the integrity of the electropherogram to calculate the RNA Integrity Number (RIN) with a value of 1–10, with 10 being the highest. The RIN (given by electropherogram) for total RNA input should be >7.

Small RNAs are purified using polyacrylamide gel electrophoresis (PAGE) (*see* Fig. 1a). A ladder containing two DNA oligonucleotides (18 bases and 36 bases) indicates the area that has to be cut for small RNA purification (*see* Fig. 1b). The subsequent cDNA library preparation is a four-step process starting with the ligation of a DNA oligonucleotide (3′-adapter) to the 3′-end of the selected RNAs, followed by ligation of a RNA oligonucleotide (5′-adapter) to the 5′-end of the selected RNAs (*see* Fig. 2). The resulting molecules are transcribed via reverse transcription (RT) and amplified via PCR.

The first ligation step at the 3′-OH of small RNAs utilizes the advantage of the enzymatic activity of T4 RNA ligase 2-truncated K227Q. The truncated ligase is unable to adenylate the 5′-end of substrates. As a result it cannot ligate the phosphorylated 3′- and 5′-ends that would result in circular RNAs or RNA concatemers. The second step of ligation at the 5′-end of the RNA is achieved by the T4 RNA ligase 1 in the presence of ATP. Within the next steps RNA is converted into cDNA by reverse transcription and PCR. The cDNA is then sequenced on a sequencing platform. Due to the fact that most high-throughput sequencers produce many millions of sequence reads in a single reaction, it is desirable to pool ("multiplex") libraries from multiple experiments into a single

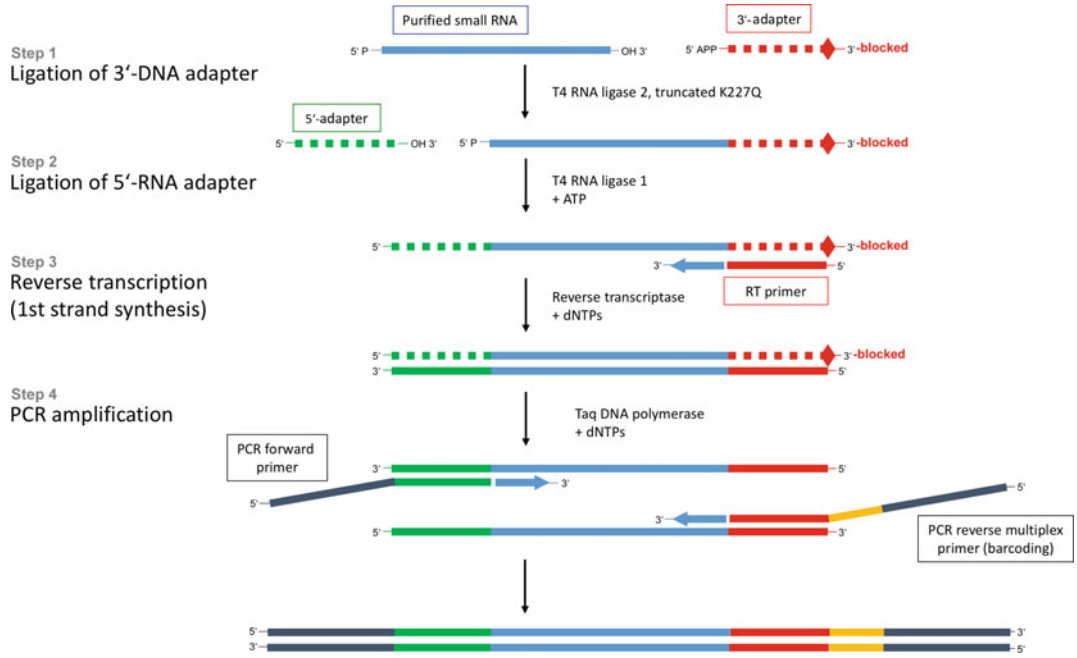
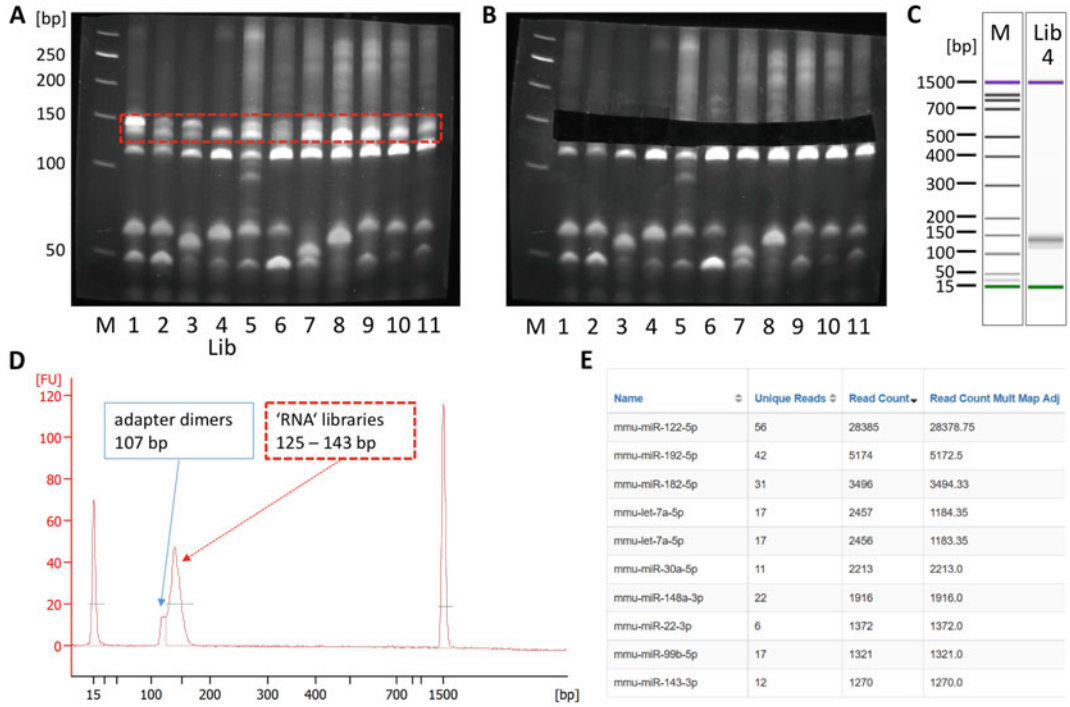


Fig. 2 Workflow for library preparation from purified small RNA fragments using adapter ligation, reverse transcription, and PCR amplification. Purified small RNAs are directly used for the construction of sequencing libraries in four steps. Hereby, the use of different barcodes for each library allows to perform multiplexed analyses of different pooled samples in a single sequencer lane. *Step 1*: Ligation of DNA adapter to the 3′-end of the RNA; *Step 2*: Ligation of adapter to the 5′-end of RNAs; *Step 3*: cDNA library synthesis by reverse transcriptase; *Step 4*: Amplification of the cDNA library

sequencing reaction and save costs. For example, a single flowcell lane from an Illumina HiSeq 2000 instrument routinely yields 150–400 million sequences, a good coverage for a single library requires only 5–10 million reads. To identify from which sample a given sequence derives a short sequence of usually at least 6 nt (index or "barcode") is incorporated into each DNA fragment during the PCR step of library preparation. At this point up to 48 index/samples/expression profiles can be processed at the same time per sequencing lane, each sample having a different index. The cDNA libraries are purified using polyacrylamide gel electrophoresis (*see* Fig. 3a). This step separates the libraries (125–143 bp) from adapter dimers (107 bp) which are simultaneously produced upon PCRs (*see* Fig. 3b). Note that depending on the relative ratio of "RNA" libraries and adapter dimers the former can still be contaminated by the latter. Contamination can be evaluated via an Agilent Bioanalyzer (*see* Fig. 3c, d). The cDNA libraries have to be mixed in an equimolar ratio for sequencing. For that concentration has to be evaluated.



Name	Unique Reads	Read Count	Read Count Mult Map Adj
mmu-miR-122-5p	56	28385	28378.75
mmu-miR-192-5p	42	5174	5172.5
mmu-miR-182-5p	31	3496	3494.33
mmu-let-7a-5p	17	2457	1184.35
mmu-let-7a-5p	17	2456	1183.35
mmu-miR-30a-5p	11	2213	2213.0
mmu-miR-148a-3p	22	1916	1916.0
mmu-miR-22-3p	6	1372	1372.0
mmu-miR-99b-5p	17	1321	1321.0
mmu-miR-143-3p	12	1270	1270.0

Fig. 3 Stringent monitoring of the workflow leads to a very high success rate of library preparation. (**a**) Polyacrylamide gel electrophoresis was performed for the purification of "RNA" libraries from tissues (here: murine liver). One discrete band corresponding to the adapter dimer (107 bp) and one or two bands corresponding to uni- or bimodal "RNA" libraries (125–143 bp) are visible. Bimodal libraries consist of two or more bands, which might be due to different sizes of small RNAs in the samples. (**b**) Gel areas containing "RNA" libraries are cut between 125 and 150 bp, each sample separately. (**c**, **d**) Quality assessment of isolated "RNA" library (i.e., "Lib" corresponding to lane 4 in Fig. **3 (a)** + **(b)**) through microcapillary electrophoresis, i.e., DNA 1000 chip on an Agilent Bioanalyzer device. (**c**) Virtual gel. (**d**) Electropherogram showing a library between 125 and 143 bp and few adapter dimers of 107 bp. When two or more positive library bands were seen on the gel, more than one peak would be expected here. (**e**) Exemplary read count table of mapped mature miRNA sequences (i.e., "Lib" from (**c**)) obtained after sequencing. These data can be obtained by few steps using sRNAtoolbox (http:/bioinfo5.ugr.es/srnatoolbox/index), directly after the packed FASTQ files were uploaded to the server. Here, the top ten mature miRNAs (read counts) are listed in descending order

1.3 Bioinformatics Analyses

After sequencing, each read needs to be assigned to the sample it comes from using the index sequence. This step is achieved by adequate software (i.e., bcl2fastq2 Conversion Software for Illumina platforms). In addition, a common filtering step is to discard or trim reads containing low quality bases. The output from the sequencer is translated to base-call quality, which depends on the sequencing platform and the version of base-calling software. The sequencing reads and the corresponding base-call qualities are delivered to the user typically as a FASTQ file (which has the extension ".fastq" or ".fq"). These FASTQ files are simple text-files containing a four-line record for each read, including its nucleotide sequence, a "+" sign separator (optionally with the read

identifier repeated), and a corresponding ASCII string of quality characters [22]. Remaining adapters are then trimmed off the reads.

For non-experts unexperienced in computational biology, an entry point for further analyses, normalization, and visualization of sequencing data can be the use of several online tools available. Those tools offer the possibility of easy learning using standard protocols, but have some limitations with respect to the customizability of experimental parameters. To convert, filter, and sort read counts from FASTQ files, Galaxy (http://usegalaxy.org/) is a very useful and helpful tool. The obtained read counts from miRNA data need to be mapped to the reference genome or compared to known mature miRNAs in the databank using, for example, sRNAtoolbox (http://bioinfo5.ugr.es/srnatoolbox) or miRBase (http://www.mirbase.org/) as a reference library (*see* Fig. 3e) [23]. Normalizing read counts by coverage is done in a variety of ways using differential expression software [24–26]. Most efficient normalization used DESeq procedures. There are tutorials available to get deeper insight into this normalization procedure (http://cgrlucb.wikispaces.com/Spring+2012+DESeq+Tutorial). MicroScope (http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1260-x) can be used for visualization and clustering of expression data. Most of these online tools provide interfaces to the most important genome databases useful for data mining.

Since new sequencing technologies and new sequencing chemistries are being developed at rapid pace, a number of sequencing platforms are widely available, for the Illumina, Ion Torrent, and PacBio technologies. Moreover the first portable nanopore sequencing devices are now being marketed by Oxford Nanopore. The here described approach with the combined power of Illumina sequencing allows a large number of samples and/or physiological conditions to be analyzed at a very affordable cost level per sample.

2 Materials

2.1 Purification of Total RNA

2.1.1 From Fresh or Frozen Tissue

1. Tissue disruption:
   (a) Cellcrusher™ and liquid nitrogen.
   (b) Ceramic Bead Tubes and PreCellys® 24/Cryloys homogenizer.
2. Enzymatic digestion and MACS cell separation for cell type purification.
3. Qiazol Lysis Reagent (Qiagen). Stored at room temperature.
4. Chloroform.
5. Isopropanol 100%.
6. RNase-DNase-free water.

2.1.2 From Cell Culture

1. Cell scraper.
2. Qiazol Lysis Reagent (Qiagen). Stored at room temperature.
3. Chloroform.
4. Isopropanol 100%.
5. RNase-DNase-free water.

2.1.3 From FFPE-Samples

1. Xylene.
2. miRNeasy FFPE Kit (Qiagen).
3. Qiazol Lysis Reagent (Qiagen). Stored at room temperature.
4. Chloroform.
5. Isopropanol 100%.
6. RNase-DNase-free water.

2.1.4 From Blood Samples

1. QIAamp RNA Blood Mini Kit (Qiagen).
2. Qiazol Lysis Reagent (Qiagen). Stored at room temperature.
3. Chloroform.
4. Isopropanol 100%.
5. RNase-DNase-free water.

2.2 Selection of Small RNAs

1. Formamide.
2. 30% Acrylamide/Bis solution (29:1). Stored at 4 °C.
3. Urea.
4. 10% Ammonium persulfate (APS) solution. Stored at −20 °C (*see* **Note 1**).
5. *N*,*N*,*N*′,*N*′-Tetramethyl-ethylenediamine (TEMED). Stored at 4 °C (*see* **Note 2**).
6. TBE buffer: 90 mM Tris, 90 mM boric acid, 10 mM EDTA.
7. 6× Loading dye: 50% (v/v) glycerol, 200 mM Tris–HCl pH 8.0, 100 mM acetic acid, 5 mM EDTA, bromophenol blue
8. Marker Oligonucleotides (100 μM) (18-mer: 5′-CAGTGGCTG GTTGAGATA-3′), (36-mer: 5′-CAGTGGCTGGTTGAGATA GTTGCTACCCCTTTCCTT-3′).
9. Equipment: Vertical electrophoresis unit 15 × 24 × 0.75 cm, spacers and comb of 0.75 cm width.
10. 24-Well cell culture plate.
11. 0.4 M NaCl solution.
12. Glycogene 20 mg/mL.
13. Ethanol 100%.
14. RNase-DNase-free water.

2.3 Construction of cDNA Libraries

2.3.1 Ligation of 3′-Adapter (Small RNA)

1. DNA 3′-adapter oligonucleotide (RA3) (0.25 μM) (5′-APP-TGGAATTCTCGGGTGCCAAGG-blocked-3′). Modifications: Adenylation-5′ (rApp) and blocking residue (Spacer C3).
2. 50% PEG 8000 solution. Stored at room temperature.
3. 10× Buffer T4 RNA Ligase 2, truncated K227Q (without ATP).
4. T4 RNA Ligase 2, truncated K227Q (240 units/μL).
5. 0.2 mL PCR tube.

2.3.2 Ligation of 5′ Adapter (Small RNA—3′ Adapter)

1. RNA 5′-adapter oligonucleotide (RA5) (1 μM) (5′-GUUCA-GAGUUCUACAGUCCGACGAUC-OH-3′).
2. 10 mM ATP.
3. T4 RNA Ligase 1 (ssRNA Ligase) (15 units/μL).

2.3.3 Reverse Transcription (5′ Adapter—Small RNA—3′ Adapter)

1. DNA Reverse transcription primer oligonucleotide (RTP) (100 μM) (5′-GCCTTGGCACCCGAGAATTCCA-OH-3′).
2. dNTPs (10 mM each).
3. 0.1 M Dithiothreitol (DTT).
4. 5× buffer SuperScript III (Thermo Fisher Scientific).
5. SuperScript III Reverse Transcriptase (130 units/μL) (Thermo Fisher Scientific).

2.3.4 Amplification and Multiplexing (PCR)

1. DNA PCR primer oligonucleotide (RP1) (100 μM) (5′-AAT-GATACGGCGACCACCGACAGGTTCAGAGTTCTA-CAGTCCGA-OH-3′).
2. DNA PCR primer barcoded oligonucleotide (RPI 01–RPI 48) (100 μM) (5′-CAAGCAGAAGACGGCATACGAGAT-NNNNNN-GTGACTGGAGTTCCTTGGCACCCGA-GAATTCCA-OH-3′). The NNNNNN sequence has to be replaced for each primer by the following hexamers (Index RPI 01–RPI 48):

Index	Sequence	Index	Sequence	Index	Sequence
RPI01	CGTGAT	RPI17	CTCTAC	RPI33	CGCCTG
RPI02	ACATCG	RPI18	GCGGAC	RPI34	GCCATG
RPI03	GCCTAA	RPI19	TTTCAC	RPI35	AAAATG
RPI04	TGGTCA	RPI20	GGCCAC	RPI36	TGTTGG
RPI05	CACTGT	RPI21	CGAAAC	RPI37	ATTCCG
RPI06	ATTGGC	RPI22	CGTACG	RPI38	AGCTAG
RPI07	GATCTG	RPI23	CCACTC	RPI39	GTATAG
RPI08	TCAAGT	RPI24	GCTACC	RPI40	TCTGAG
RPI09	CTGATC	RPI25	ATCAGT	RPI41	GTCGTC

RPI10	AAGCTA	RPI26	GCTCAT	RPI42	CGATTA
RPI11	GTAGCC	RPI27	AGGAAT	RPI43	GCTGTA
RPI12	TACAAG	RPI28	CTTTTG	RPI44	ATTATA
RPI13	TTGACT	RPI29	TAGTTG	RPI45	GAATGA
RPI14	GGAACT	RPI30	CCGGTG	RPI46	TCGGGA
RPI15	TGACAT	RPI31	ATCGTG	RPI47	CTTCGA
RPI16	GGACGG	RPI32	TGAGTG	RPI48	TGCCGA

3. 2× Phusion Hot Start Flex Master Mix (NEB).
4. 3 M sodium acetate.
5. Ethanol 100%.
6. RNase-DNase-free water.

2.3.5 Selection of Libraries

1. 30% Acrylamide/Bis solution (29:1). Stored at 4 °C.
2. 10% Ammonium persulfate (APS) solution. Stored at −20 °C (*see* **Note 1**).
3. *N*,*N*,*N*′,*N*′-Tetramethyl-ethylenediamine (TEMED). Stored at 4 °C (*see* **Note 2**).
4. TBE buffer: 90 mM Tris, 90 mM boric acid, 10 mM EDTA.
5. 6× Loading dye: 50% (v/v) glycerol, 200 mM Tris–HCl pH 8.0, 100 mM acetic acid, 5 mM EDTA, bromophenol blue.
6. 50 bp Ladder.
7. Equipment: Vertical electrophoresis unit 15 × 24 × 0.75 cm, spacers and comb of 0.75 cm width.
8. 24-Well cell culture plate.
9. 0.4 M NaCl solution.
10. Ethanol 100%.
11. RNase-DNase-free water.

2.4 Multiplex Sequencing and Data Analyses

1. Illumina HiSeq or MiSeq platforms.
2. Bioinformatic statistical analyses and data visualization tools.

3 Methods

3.1 Purification of Total RNA

3.1.1 From Fresh or Frozen Tissue

1. Tissue disruption:
   (a) Disrupt tissue using liquid nitrogen and a pre-chilled Cellcrusher™. Briefly transfer the radically grinded tissue to a microcentrifuge tube, add 1 mL Qiazol Lysis Reagent and vortex. Add 200 µL chloroform, vortex, and incubate for 5 min. Centrifuge at 12,000 × *g* for 10 min.

(b) Add Ceramic Bead Tubes (Qiagen) with fresh or frozen tissue (< 300 mg) and 700 μL of Qiazol Lysis Reagent and immediately homogenize using a PreCellys® 24/Cryloys homogenizer for 1–2 pulses of 20 s at room temperature. Add 200 μL chloroform, vortex, and incubate for 5 min. Centrifuge at 12,000 × *g* for 10 min.

2. Cell type purification:
   (a) Use enzymatic digestion and MACS cell separation according to the manufacturer's instructions for cell type purification. Briefly transfer cells to a microcentrifuge tube, add 1 mL Qiazol Lysis Reagent, and vortex. Add 200 μL chloroform, vortex, and incubate for 5 min. Centrifuge at 12,000 × *g* for 10 min.
3. Upon centrifugation three phases are visible, an organic (pink) phase at the bottom of the tube, an aqueous clear phase on top of it, and a slightly white interphase between those two.
4. Carefully transfer the aqueous phase that contains RNA to a fresh microcentrifuge tube (*see* **Note 3**).
5. Measure the aqueous volume, add one volume of isopropanol 100%, and vortex.
6. Centrifuge at 12,000 × *g* at 4 °C for 30 min and wash pellet twice with ethanol 70%.
7. Air-dry the pellet.
8. Resuspend pellet in 20 μL of RNase-DNase-free water (*see*-**Note 4**).
9. Concentration can be measured using a fluorimeter and RNA integrity can be verified using a Bioanalyzer (*see* **Note 5**).

3.1.2 From Cell Culture

1. Use a cell scraper to detach cells from the bottom of a cell culture flask and transfer to a microcentrifuge tube.
2. Centrifuge at 300 × *g* for 5 min and remove the supernatant.
3. Briefly add 1 mL Qiazol Lysis Reagent to the pellet and transfer into a microcentrifuge tube. Vortex and add 200 μL chloroform, vortex, and incubate for 5 min. Centrifuge at 12,000 × *g* for 10 min.
4. Upon centrifugation three phases are visible, an organic (pink) phase at the bottom of the tube, an aqueous clear phase on top of it, and a slightly white interphase between those two.
5. Carefully transfer the aqueous phase that contains RNA to a fresh microcentrifuge tube (*see* **Note 3**).
6. Measure the aqueous volume, add one volume of isopropanol 100%, and vortex.
7. Centrifuge at 12,000 × *g* at 4 °C for 30 min and wash pellet twice with ethanol 70%.

8. Air-dry pellet.
9. Resuspend pellet in 20 μL of RNase-DNase-free water (*see* **Note 4**).
10. Concentration can be measured using a fluorimeter and RNA integrity can be verified using a Bioanalyzer, respectively (*see* **Note 5**).

3.1.3 From FFPE-Samples

1. Xylene is used to get rid of paraffin from the formalin-fixed samples.
2. Isolate RNA using the miRNeasy FFPE Kit according to the manufacturer's instructions.
3. Resuspend RNA in 10 μL of RNase-DNase-free water (*see* **Note 4**).
4. Concentration can be measured using a fluorimeter and RNA integrity can be verified using a Bioanalyzer (*see* **Note 5**).

3.1.4 From Blood-Samples

1. Isolate RNA using the QIAamp RNA Blood Mini Kit according to the manufacturers' instructions.
2. Resuspend RNA in 10 μL of RNase-DNase-free water (*see* **Note 4**).

3.2 Selection of Small RNAs (17% Polyacrylamide Gel Electrophoresis)

1. Dissolve 17 g of urea in 20 mL of acrylamide solution (30%), 4 mL of TBE buffer (5×), and 7 mL of water. Add 160 μL of ammonium persulfate (10%) and 80 μL of TEMED and cast gel within a 15 × 24 × 0.75 cm gel cassette.
2. Immediately insert a 10- or 12-well gel comb without introducing air bubbles.
3. Add one volume of formamide to each RNA sample (*see* **Note 6**).
4. Mix a ladder sample with 1 μL of each marker oligonucleotide (18-mer, 36-mer) and 12 μL RNase-DNase-free water and add 14 μL of formamide.
5. Heat RNA and ladder samples for 2 min at 70 °C to minimize secondary structures. Store on ice.
6. Add 5 μL of loading dye (6×) to each sample.
7. Rinse gel wells. Be aware of loading the marker sample on one side of the gel, not in an internal well, for easy visualization (*see* Fig. 1).
8. Carry out electrophoresis starting with 100 V for 15 min, then 350 V for 2.5 h.
9. Following electrophoresis (*see* **Note 7**), separate the gel plates with the use of a spatula. Make sure that the gel remains on one of the glass plates. Carefully transfer the gel to a water bath containing 0.5 μg/mL ethidium bromide. Incubate for 10 min with gentle shaking.

10. Transfer gel to a plastic film and place it on a UV absorbing plate (*see* **Note 7**).
11. Visualize the ladder and, in case of RNAs purified from cells or tissues, the first 5 cm of migration of each sample (*see* **Notes 8** and **9**) (*see* Fig. 1a).
12. Cut between the marker bands using a scalpel, one band per track. Cut gel fragments between 18 and 36 bases and put them in a 24-well cell culture plate. Each sample in a separate well (*see* Fig. 1b).
13. Add 600 μL of 0.4 M NaCl and incubate over night at 4 °C with gentle shaking.
14. Transfer 400 μL to a 1.5 mL microcentrifuge tube and add 2.5 volumes (1 mL) of ethanol 100% and 1 μL of glycogen (20 mg/mL) (*see* **Note 10**).
15. Centrifuge at 12,000 × *g* at 4 °C for 30 min and wash pellet twice with ethanol 70%.
16. Resuspend pellet in 10 μL of RNase-DNase-free water (*see* **Note 11**).

3.3 Construction of cDNA Libraries

3.3.1 Ligation of 3′ Adapter (Small RNA)

1. Small RNAs in 10 μL of RNase-DNase-free water in a 0.2 mL PCR tube.
2. Add 1 μL of 5′-adenylated 3′-adapter oligonucleotide (RA3) (0.25 μM).
3. Heat samples at 70 °C for 2 min and store on ice.
4. Add 4.8 μL of PEG 8000 50% (*see* **Note 12**).
5. Add 2.2 μL of buffer T4 RNA Ligase 2 truncated K227Q (without ATP) and 0.8 μL of T4 RNA Ligase 2, truncated K227Q (200 units/μL). When working with multiple samples make a master mix of T4 RNA Ligase 2 enzyme and buffer (up-sized by 10%) and distribute.
6. Incubate at 25 °C for 90 min (*see* **Note 13**).

3.3.2 Ligation of 5′ Adapter (Small RNA—3′ Adapter)

1. Add 1 μL of 5′-adapter oligonucleotide (RA5) (1 μM).
2. Heat samples at 70 °C for 2 min and store on ice.
3. Add 1 μL of 10 mM ATP and 0.75 μL of T4 RNA Ligase 1 (15 units/μL). When working with multiple samples produce a master mix of T4 RNA Ligase 1 enzyme and ATP (up-sized by 10%) and distribute.
4. Incubate at 25 °C for 90 min (*see* **Note 13**).

3.3.3 Reverse Transcription (5′ Adapter—Small RNA—3′ Adapter)

1. Add 0.5 μL of reverse transcription primer oligonucleotide (RTP) (100 μM).
2. Heat samples at 70 °C for 2 min and store on ice.

3. Add 1.5 μL of dNTPs (10 mM each), 1.5 μL of 0.1 M Dithiothreitol (DTT), 6 μL of 5× Buffer SuperScript III and 0.65 μL of SuperScript III Reverse Transcriptase (130 units/μL). When working with multiple samples produce a master mix (up-sized by 10%) and distribute.
4. Incubate at 50 °C for 90 min (*see* **Note 13**).

3.3.4 Amplification and Multiplexing (PCR)

1. Add 0.6 μL of RNA PCR primer oligonucleotide (RP1) (100 μM), 30 μL of Phusion Hot Start Flex 2× Master Mix and 23 μL of DNase-RNase-free water. When working with multiple samples produce a master mix (up-sized by 10%) and distribute.
2. Add 1.2 μL of RNA PCR primer barcoded oligonucleotide (i.e., RPI 01 which specific bases are underlined in the sequence below) (100 μM) (5′-CAAGCAGAAGACGGCA-TACGAGAT-CGTGAT-GTGACTGGAGTTCCTTGG-CACCCGAGAATTCCA-OH-3′) to each sample. Each sample having a different barcoded oligonucleotide.
3. Split each sample in four tubes with equal volume to enhance thermic exchanges and optimize PCR efficiency.
4. Perform PCR with the following conditions:

Initial denaturation	**98 °C**	**1 min**	**1×**
Denaturation	98 °C	20 s	16–20× (*see* **Note 14**)
Annealing	55 °C	30 s	
Elongation	72 °C	25 s	
Final elongation	72 °C	10 min	1×
End	25 °C	∞	1×

5. Pool PCR products of the four tubes for each sample. Add 6 μL of 3 M sodium acetate and 150 μL of ethanol 100% to each reaction.
6. Centrifuge at 12,000 × *g* at 4 °C for 30 min and wash pellet twice with ethanol 70%.
7. Resuspend pellet in 20 μL of RNase-DNase-free water (*see* **Note 13**).

3.4 Selection of cDNA Libraries (6% Polyacrylamide Gel Electrophoresis)

1. Mix 8 mL of acrylamide mixture (30%), 4 mL of TBE buffer (5×) and 28 mL of water. Add 160 μL of ammonium persulfate (10%) and 80 μL of TEMED and cast gel within a 15 × 24 × 0.75 cm gel cassette.
2. Immediately insert a 10–12-well gel comb without introducing air bubbles.

3. Prepare the ladder sample with 1 μL of 50 bp Ladder added to 19 μL of RNase-DNase-free water.
4. Add 5 μL of loading dye (6×) to library and ladder samples.
5. Carry out electrophoresis starting with 100 V for 15 min, then 250 V for 3.5 h.
6. Following electrophoresis (*see* **Note 7**), cool gel plates with ice for 10 min. Separate the gel plates with the use of a spatula. The gel remains on one of the glass plates. Carefully transfer to a water bath containing ethidium bromide (0.5 μg/mL). Incubate for 10 min with gentle shaking.
7. Transfer gel to a plastic film and place it on a UV absorbing plate (*see* **Note 7**).
8. Visualize the entire gel (*see* **Note 9**). Two bands >100 bp are visible (*see* **Note 15**). One at 107 bp (side reaction, adapter dimers), the other between 125 and 143 bp ("RNA" library) (*see* Fig. 3a).
9. Cut gel fragments between 125 and 150 bp and put them in a 24-well cell culture plate. Each sample in a separate well (*see* Fig. 3b).
10. Add 600 μL of 0.4 M NaCl and incubate over night at 4 °C with gently shaking.
11. Transfer 400 μL of the eluate to a 1.5 mL microcentrifuge tube and add 1.5 volumes (1 mL) of ethanol 100%.
12. Centrifuge at 12,000 × *g* at 4 °C for 30 min and wash pellet twice with ethanol 80%.
13. Resuspend pellet in 10 μL of RNase-DNase-free water.
14. Concentration and quality of the libraries can be measured using a fluorimeter or a Bioanalyzer (*see* Fig. 3c, d).

3.5 Multiplex Sequencing

1. Sequencing of the multiplexed cDNA libraries is performed on Illumina HiSeq 2000 using single read flowcells or Illumina MiSeq or NextSeq platforms using double read (paired-end) flowcells, which can be run in single or double end mode.
2. Sequencing data needs to be adapter trimmed and further analyzed using bioinformatics statistic programs and data visualization tools, i.e., Galaxy (http://usegalaxy.org/), sRNA-toolbox (http://bioinfo5.ugr.es/srnatoolbox) (*see* Fig. 3e).

4 Notes

1. APS loses its catalytic power when left at higher temperatures.
2. Open and pipette in a hood.
3. Be really careful transferring the aqueous phase. Avoid disruption of the interphase (mainly containing genomic

DNA), to be sure not to contaminate the aqueous phase during pipetting.

4. Wear gloves and work in a clean workspace. Quickly process the RNA and keep on ice when possible. Long time storage at −80 °C.
5. Do not measure the OD with a Photometer or Fluorimeter (i.e., *Quantus* (Promega)) in case of low RNA concentration (50–100 ng/μL). The OD will not reflect the RNA concentration due to contamination of absorbent components. If possible use a Bioanalyzer instead.
6. The amount of total RNA used for the isolation of small RNAs can be scaled up to 8 μg per well. Be careful not to overload the gel, to assure optimal separation of the different sizes of RNA fragments during electrophoresis.
7. The gel is very fragile, it easily rips to shreds.
8. Be sure to cover the area of small RNAs in the gel very well, when visualizing on the UV-projector. Alternatively, gel can be visualized on a non-UV transilluminator (i.e., UVIBlue transilluminator (Uvitec Cambridge) (http://www.uvitec.co.uk/products/uviblue.html)). This allows visualization without any risk of damaging the RNA.
9. Minimize time of exposure to UV-light to reduce single-strand breaks. UV exposure dramatically reduces library sequencing efficiency.
10. Glycogen is used as a carrier.
11. The quantity of purified small RNAs is usually too low to allow any concentration to be measured. Purified RNAs obtained from 0.2 to 1 μg of total RNAs however provide abundant cDNA libraries.
12. PEG 8000 50% is very viscous. Keep it at room temperature and pipette very slowly and cautiously. Proceed tube by tube.
13. Protocol can be paused after this step with sample storage at −20 °C.
14. PCR cycles can be scaled up to 20 when using low quantities of total RNA for small RNA purification. Do not go further as more cycles result in less PCR products.
15. The second ligation step leads to a side reaction and the formation of 5′-adapter-3′-adapter products. This reaction can be quenched by the addition of the RT-primer before the second ligation step. Add 0.5 μL reverse transcription primer oligonucleotide (RTP) (100 μM) after the first ligation step and heat samples at 70 °C for 2 min and store on ice. Incubate at 25 °C for 30 min. Proceed to the second ligation step as described, but never heat samples again.

References

1. Fire AZ (2007) Gene silencing by double-stranded RNA. Cell Death Differ 14 (12):1998–2012. doi:10.1038/sj.cdd.4402253
2. Kanasty R, Dorkin JR, Vegas A, Anderson D (2013) Delivery materials for siRNA therapeutics. Nat Mater 12(11):967–977. doi:10.1038/nmat3765
3. Burnett JC, Rossi JJ (2012) RNA-based therapeutics: current progress and future prospects. Chem Biol 19(1):60–71. doi:10.1016/j.chembiol.2011.12.008
4. Sullenger BA, Nair S (2016) From the RNA world to the clinic. Science 352 (6292):1417–1420. doi:10.1126/science.aad8709
5. Cech TR, Steitz JA (2014) The noncoding RNA revolution-trashing old rules to forge new ones. Cell 157(1):77–94. doi:10.1016/j.cell.2014.03.008
6. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR (2005) MicroRNA expression profiles classify human cancers. Nature 435(7043):834–838. doi:10.1038/nature03702
7. Marcucci G, Radmacher MD, Maharry K, Mrozek K, Ruppert AS, Paschka P, Vukosavljevic T, Whitman SP, Baldus CD, Langer C, Liu CG, Carroll AJ, Powell BL, Garzon R, Croce CM, Kolitz JE, Caligiuri MA, Larson RA, Bloomfield CD (2008) MicroRNA expression in cytogenetically normal acute myeloid leukemia. N Engl J Med 358(18):1919–1928. doi:10.1056/NEJMoa074256
8. Pfeffer S, Zavolan M, Grasser FA, Chien M, Russo JJ, Ju J, John B, Enright AJ, Marks D, Sander C, Tuschl T (2004) Identification of virus-encoded microRNAs. Science 304 (5671):734–736. doi:10.1126/science.1096781
9. Moore BT, Xiao P (2013) MiRNAs in bone diseases. Microrna 2(1):20–31
10. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T (2001) Identification of novel genes coding for small expressed RNAs. Science 294 (5543):853–858. doi:10.1126/science.1064921
11. Liu X, Luo G, Bai X, Wang XJ (2009) Bioinformatic analysis of microRNA biogenesis and function related proteins in eleven animal genomes. J Genet Genomics 36(10):591–601. doi:10.1016/S1673-8527(08)60151-4
12. Seitz H, Zamore PD (2006) Rethinking the microprocessor. Cell 125(5):827–829. doi:10.1016/j.cell.2006.05.018
13. Du T, Zamore PD (2005) microPrimer: the biogenesis and function of microRNA. Development 132(21):4645–4652. doi:10.1242/dev.02070
14. Huang Y, Shen XJ, Zou Q, Zhao QL (2010) Biological functions of microRNAs. Bioorg Khim 36(6):747–752
15. Swarts DC, Makarova K, Wang Y, Nakanishi K, Ketting RF, Koonin EV, Patel DJ, van der Oost J (2014) The evolutionary journey of Argonaute proteins. Nat Struct Mol Biol 21 (9):743–753. doi:10.1038/nsmb.2879
16. Brodersen P, Voinnet O (2009) Revisiting the principles of microRNA target recognition and mode of action. Nat Rev Mol Cell Biol 10 (2):141–148. doi:10.1038/nrm2619
17. Brennecke J, Stark A, Russell RB, Cohen SM (2005) Principles of microRNA-target recognition. PLoS Biol 3(3):e85. doi:10.1371/journal.pbio.0030085
18. Karginov FV, Conaco C, Xuan Z, Schmidt BH, Parker JS, Mandel G, Hannon GJ (2007) A biochemical approach to identifying microRNA targets. Proc Natl Acad Sci U S A 104 (49):19291–19296. doi:10.1073/pnas.0709971104
19. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 433 (7027):769–773. doi:10.1038/nature03315
20. Vigneault F, Sismour AM, Church GM (2008) Efficient microRNA capture and bar-coding via enzymatic oligonucleotide adenylation. Nat Methods 5(9):777–779. doi:10.1038/nmeth.1244
21. Baroin-Tourancheau A, Benigni X, Doubi-Kadmiri S, Taouis M, Amar L (2016) Lessons from microRNA sequencing using illumina technology. Adv Biosci Biotechnol 7:319–328
22. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38 (6):1767–1771. doi:10.1093/nar/gkp1137
23. Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39 (Database issue):D152–D157. doi:10.1093/nar/gkq1027

24. Oshlack A, Robinson MD, Young MD (2010) From RNA-seq reads to differential expression results. Genome Biol 11(12):220. doi:10.1186/gb-2010-11-12-220
25. Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinform 11:94. doi:10.1186/1471-2105-11-94
26. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11(3):R25. doi:10.1186/gb-2010-11-3-r25

Chapter 13

MicroRNA Expression Analysis Using Small RNA Sequencing Discovery and RT-qPCR-Based Validation

Alan Van Goethem, Pieter Mestdagh, Tom Van Maerken, and Jo Vandesompele

Abstract

miRNAs are small noncoding RNA molecules that function as regulators of gene expression. Deregulated miRNA expression has been reported in various diseases including cancer. Due to their small size and high degree of homology, accurate quantification of miRNA expression is technically challenging. In this chapter, we present two different technologies for miRNA quantification: small RNA sequencing and RT-qPCR.

Key words miRNA, Small RNA sequencing, RT-qPCR, miRNA annotation, Normalization

1 Introduction

MicroRNAs are a large class of small noncoding RNAs that regulate gene expression at the posttranscriptional level. To date, over 28,000 hairpin miRNAs, giving rise to more than 35,000 mature miRNAs in 223 species, have been described in the miRBase sequence database (version 21, http://www.mirbase.org), including 2588 mature human miRNAs. The prominent role of miRNAs in virtually every aspect of cell biology and their involvement in disease have led to the development of both diagnostic and prognostic miRNA expression signatures as well as miRNA-based therapeutics. As miRNAs function as biological rheostats concurrently affecting several target genes, even subtle alterations in their abundance may have substantial impact. Unfortunately, their small size, low abundance, and the high degree of homology among miRNA family members make accurate quantification of mature miRNA expression levels technically challenging. Several platforms are available for assessing miRNA abundance, based on (micro-array) hybridization, reverse transcription qPCR (RT-qPCR), or (small RNA) sequencing.

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_13,

The miRQC study has comprehensively assessed different miRNA expression platforms using quantifiable performance metrics [1]. The result is an unbiased comparison of accuracy, specificity, sensitivity, and reproducibility among 12 different platforms from 9 different vendors. Each platform was found to have its strengths and weaknesses. The study outcome should aid researchers making an informed selection of platform corresponding to the experimental setting and the specific research question.

Since the start of the miRQC study, RNA sequencing has witnessed continuing technical and workflow improvements and decreasing costs. Combined with the possibility to assess a large number of small RNAs including the discovery of previously uncharacterized miRNAs, this has rendered small RNA sequencing as the gold standard method for miRNA discovery and quantification. Typically, and as recommended in the miRQC study conclusions, the initial sequencing based screening experiment is followed by validation of obtained results using RT-qPCR.

1.1 Small RNA Sequencing

Similar to most RNA sequencing approaches, small RNA sequencing requires the construction of cDNA libraries (Fig. 1). The initial step of library preparation is adapter ligation. The adapters serve as a template for primer-based RT, amplification, and sequencing. All RNA molecules containing a 5′ phosphate and 3′ hydroxyl group will be subjected to both 5′ and 3′ single stranded RNA adapter ligation. Adapter ligation is followed by reverse transcription of the adapter-ligated RNA into cDNA and PCR amplification of the cDNA libraries. During the PCR amplification step, each library is tagged with a unique index that enables identification of the library origin of individual reads when analyzing sequencing data, thus making it possible to simultaneously sequence a few dozen samples. Library preparation kits from different vendors are available for the preparation of small RNA libraries, they mainly differ in the process of adapter ligation and the presence or absence of adapter dimer removal [2, 3]. After PCR amplification, a library size selection step is performed to selectively enrich and select for the miRNA-containing fraction of the resulting libraries. Size selection involves a size-based separation of the library by agarose gel electrophoresis followed by DNA staining and the collection of the band containing the miRNA fraction. This can be performed manually or, alternatively, through the use of fully automated size-selection systems.

1.2 Processing Small RNA Sequencing Data

High-throughput sequencing of small RNAs leads to the generation of considerable amounts of data. The processing of these data for expression analysis can be roughly divided into two steps: preprocessing of raw sequencing data into a miRNA count table and differential gene expression analysis, including normalization of miRNA count data followed by statistical analyses. During

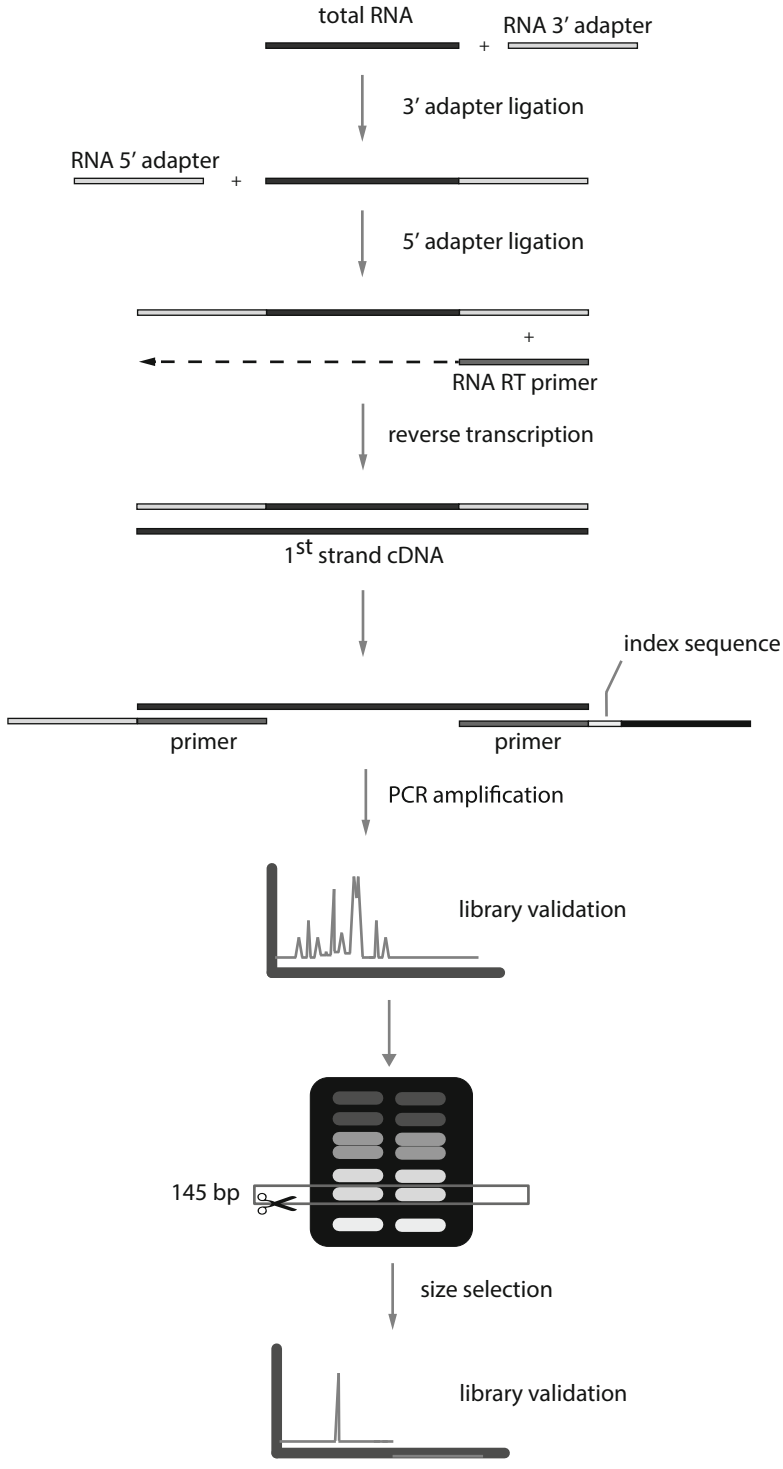


Fig. 1 Schematic overview of the TruSeq small RNA library preparation protocol

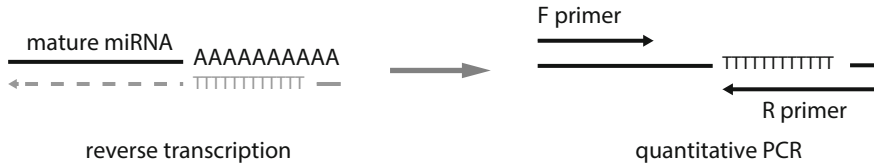


Fig. 2 Schematic overview of the universal PCR profiling platform

preprocessing adapter trimming, quality control and read mapping is performed, after which annotation information is retrieved from miRBase (and other reference databases if needed) to create a miRNA count table. As small RNA sequencing data are subject to various sources of technical variation (e.g., differences in library size or GC-content) it is necessary to perform some form of data normalization to correct for these variations. Several normalization strategies for (small) RNA-seq data have been developed (compared in [4]), but consensus on the optimal normalization method is currently lacking. The most widely used algorithm is the negative binomial-based approach DESeq2 [5] (*see* **Note 1**). DESeq2 applies the median of the ratios of each gene in one sample over the geometric mean of that gene across all samples as a scaling factor for all the genes in that sample [5].

1.3 RT-qPCR (Universal Primer)

As recommended in the miRQC study conclusions, any screening study should be followed by targeted validation using an independent method. Therefore, sequencing experiments are often followed by RT-qPCR validation of obtained results, typically a limited number of differentially expressed miRNAs. RT-qPCR-based quantification of miRNAs offers superior sensitivity and accuracy. To enable the detection of short RNA molecules like miRNAs by RT-qPCR, the reverse transcription (RT) reaction requires modification. The most widely used RT-qPCR platforms are based on either the use of stem-loop RT primers or polyadenylation of the mature miRNA to enable RT [6]. Here, we describe a protocol for universal RT-qPCR, meaning the cDNA can be used for the quantification of any miRNA. This approach is based on polyadenylation of the mature miRNA prior to oligo-dT primed cDNA synthesis (Fig. 2) [7].

1.4 Identification of Stably Expressed miRNAs for Data Normalization

To distinguish technical variation from true biological difference, it is important to perform proper normalization of the obtained data. Using the geometric mean of multiple stable reference genes is widely accepted as the gold standard for the normalization of RT-qPCR data [8]. Using well-established algorithms such as geNorm, stably expressed reference genes are typically identified out of several candidate reference genes in a small pilot study. However, when it comes to normalization of microRNA data, there are no predefined sets of candidate reference microRNAs and all too often, small

nuclear or nucleolar RNAs (such as U6, U24, and U26) are used instead. We strongly advise against these internal controls as sn(o) RNAs are transcribed from a different RNA polymerase and have different functions than miRNAs. For the measurement of a large unbiased set of miRNAs we have instead published the use of the global mean miRNA expression for accurate normalization [9, 10]. However, in the context of focused validation experiments it is not possible to rely on whole-genome based normalization strategies as one is typically interested in the validation of a limited number of differentially expressed miRNAs. For this kind of experiments it is possible to identify miRNAs that resemble the global mean expression value and whose geometric mean can be used to mimic global mean normalization [9].

1.5 miRNA Annotation

The concluding step in most miRNA studies is reporting of experimental findings. From 2002, miRBase has emerged as the reference database of miRNA nomenclature. Since then miRBase underwent numerous additions and deletions of miRNA records, adaptations into more complex naming structures and changes in annotated miRNAs. These changes are the necessary consequence of increasing insights into the (mi)RNA world but they also give rise to substantial ambiguity concerning miRNA annotation in literature. Ignoring sequence annotation changes has led to erroneous interpretation, comparison, and integration of miRNA study results (*see* **Note 2**). To resolve these issues, our lab has developed miRBaseTracker (www.mirbasetracker.org), an online tool that enables comparison of all current and historical miRNA annotation data present in miRBase [11].

2 Materials

2.1 Small RNA Sequencing

1. TruSeq Small RNA library preparation kit V2 containing 10 mM ATP, HML (Ligation buffer), RNA 3′ adapter, RNA 5′ adapter, RNase Inhibitor, Stop solution, T4 RNA ligase, 25 mM dNTP mix, PCR mix, RNA PCR Primer, RNA PCR Primer Index, RNA RT Primer, RNase Inhibitor, 5× First Strand buffer, and 100 mM DTT.
2. T4 RNA ligase 2, deletion mutant, 200 U/μL.
3. Superscript II Reverse Transcriptase, 200 U/μL.
4. Agilent High Sensitivity DNA kit containing High Sensitivity DNA chips, High Sensitivity DNA Ladder, High Sensitivity DNA Markers (35–10,380 bp), High Sensitivity DNA dye concentrate, and High Sensitivity DNA gel Matrix.
5. Agilent 2100 Bio-Analyzer.
6. Nuclease-free water.

7. Pippin Prep cassettes (3% dye-free with legacy marker H) plus loading buffer and extra running buffer.
8. Glycogen, 20 μg/μL.
9. 3 M NaOAc.
10. 100% ethanol, −20 °C.
11. 70% ethanol.
12. 10 mM Tris–HCl, pH 8.5.
13. 5 μM Library quantification primer assays, Forward Primer: AATGATACGGCGACCACCGA, Reverse Primer: CAAGCAGAAGACGGCATACGA.
14. SsoAdvanced universal SYBR Green supermix.
15. Qubit DS DNA HS assay kit.
16. NextSeq 500 Mid/High Output V2 kit, 75 cycles.

2.2 Universal Primer RT-qPCR

1. miScript II RT kit containing miScript Reverse Transcriptase Mix, 10× miScript Nucleics mix, 5× miScript HiSpec Buffer, and 5× miScript HiFlex buffer.
2. Optional: miScript PreAMP kit containing 5× miScript PreAMP Buffer, HotStarTaq DNA Polymerase (2 U/μL), miScript PreAMP Universal Primer.
3. miScript SYBR Green PCR kit containing QuantiTect SYBR Green PCR Master Mix and 10× miScript Universal Primer.
4. 5 μM miScript Primer Assay.
5. Nuclease-free water.
6. TE buffer: pH 8.0, 10 mM Tris–HCl, 1 mM EDTA (for dissolving Primer assays).

3 Methods

3.1 Small RNA Sequencing

1. Add 5 μL of RNA sample (between 0.1 and 1 μg of total RNA derived from tissues or cells) to 1 μL of RNA 3′ adapter (*see* **Note 3**). Spin to collect the liquid. It is important to keep RNA on ice at all times. Do not vortex RNA.
2. Incubate the mixture at 70 °C for 2 min and immediately place on ice.
3. Combine 2 μL of Ligation buffer (HML), 1 μL of RNase inhibitor, and 1 μL of T4 RNA ligase, deletion mutant per sample (*see* **Note 4** for the preparation of mixtures when processing multiple samples). Pipet up and down and briefly spin. Add 4 μL of the ligation mix to the reaction tube from **step 2** and incubate for 1 h at 28 °C.
4. After 1 h, quickly spin, add 1 μL of Stop solution, and continue to incubate at 28 °C for 15 min.

5. Incubate RNA 5′ adapter at 70 °C for 2 min and place immediately on ice. Prepare the 5′ adapter ligation mix by combining 1 μL of 5′ RNA adapter, 1 μL of 10 mM ATP, and 1 μL T4 RNA ligase per sample. Pipet up and down and briefly spin. Add 3 μL of ligation mix to the reaction tube from **step 3** and incubate for 1 h at 28 °C.
6. Dilute the 25 mM dNTPs twofold to a 12.5 mM mix using nuclease-free water. Add 1 μL of RNA RT primer to 6 μL of 5′-3′ ligated RNA and incubate at 70 °C for 2 min. Immediately place on ice.
7. Prepare the RT mixture by combining 2 μL of 5× First Strand Buffer, 0.5 μL of 12.5 mM dNTP mix, 1 μL of 100 mM DTT, 1 μL of RNase Inhibitor, and 1 μL SuperScript II Reverse transcriptase per sample. Pipet up and down and briefly spin. Add 5.5 μL of RT mixture to the reaction tube from **step 5**.
8. Incubate at 50 °C for 1 h and immediately place on ice.
9. Prepare the PCR amplification mix by combining 2 μL of RNA PCR primer, 25 μL of PCR mix, and 8.5 μL of nuclease-free water per sample. Add 35.5 μL of the PCR amplification mix to 12.5 μL of RT product from **step 8**. Add 2 μL of RNA PCR Primer Index.
10. Run the PCR reaction as follows: 98 °C for 30 s, (98 °C for 10 s, 60 °C for 30 s, 72 °C for 15 s) × 11 (*see* **Note 5**), 72 °C for 10 min, 4 °C hold.
11. Dilute 1 μL of each library twofold using nuclease-free water and run on a High Sensitivity DNA chip to perform quality control analysis on the library. Figure 3 shows the typical

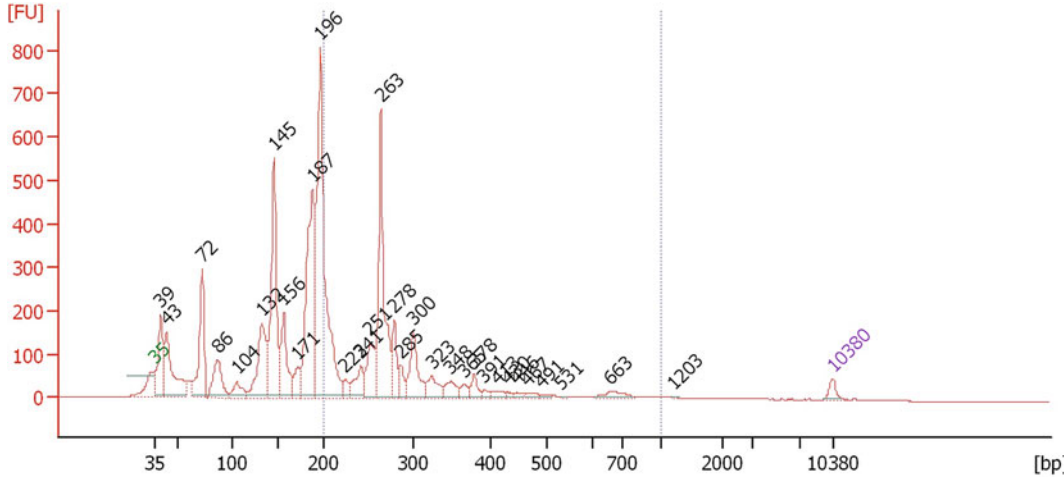


Fig. 3 Profile of a small RNA library prepared from 750 ng of tumor-derived RNA and run on an Agilent 2100 using a High Sensitivity DNA chip. The peak at 145 bp represents the miRNA fraction of the library

profile of a library prepared from 750 ng of tumor-derived RNA.

12. Load individual libraries in the lanes of a 3% agarose dye-free marker H cassette and run on a Pippin Prep with a specified collection range of 125–153 bp. Selecting this size range should maximize the collection of the miRNA fraction with minimal contaminant RNAs.
13. Collect 40 µL of resulting library from the collection well and add 2 µL of glycogen, 30 µL of 3 M sodium acetate, and 977 µL of 100% ethanol (−20 °C). Immediately centrifuge at 20,000 × *g* for 20 min at 4 °C in a fixed-angle centrifuge. Remove and discard the supernatant, leaving the pellet intact. Wash the pellet with 500 µL of 70% ethanol and centrifuge at 20,000 × *g* for 2 min at room temperature. Remove and discard the supernatant leaving the pellet intact. Dry the pellet by placing the tube, lid open, in a 37 °C heat block for 5–10 min or until dry. You will observe a shift from a white, opaque pellet to a transparent pellet if completely dry and pure. Dissolve the pellet in 20 µL 10 mM Tris–HCl, pH 8.5.
14. Dilute 1 µL of each library twofold and run on a High Sensitivity DNA chip to perform quality control analysis on the library. Figure 4 shows the typical profile of a library prepared from 750 ng of tumor-derived RNA after size selection.
15. Dilute 1 µL of each library 100,000-fold using nuclease-free water through serial dilution. Combine per sample 2.5 µL of SsoAdvanced universal SYBR Green supermix with 0.25 µL of each primer. Distribute PCR mixture in the PCR reaction and

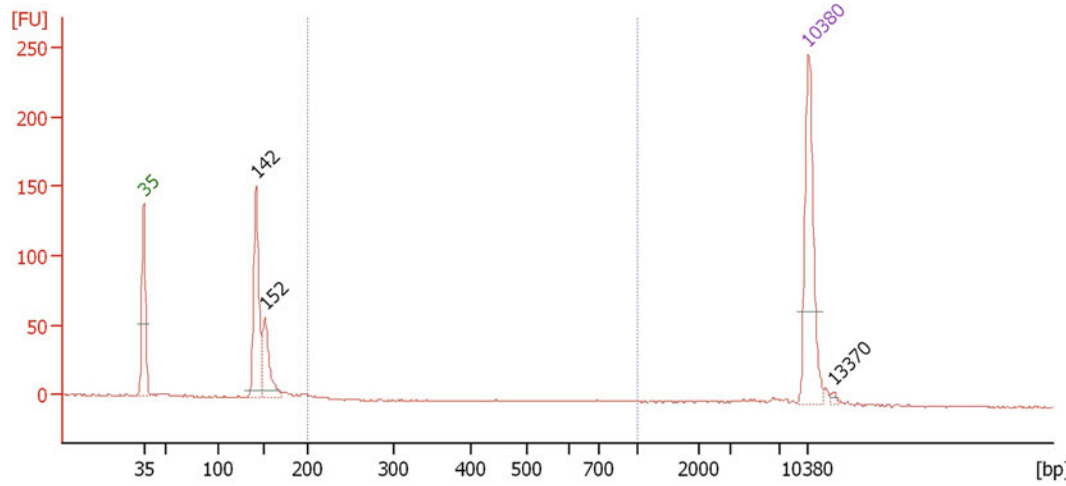


Fig. 4 Profile of a small RNA library prepared from 750 ng of tumor-derived RNA and run on an Agilent 2100 using a High Sensitivity DNA chip after library size selection. Only the library fraction containing miRNAs is retained, represented by the peak at 142 bp

add 2 μL of diluted library (triplicate reactions). Run the PCR reaction as follows: 95 °C for 15 min, (95 °C for 5 s, 60 °C for 30 s, 72 °C for 1 s) × 45 cycles, followed by melting curve analysis.

16. Prepare an equimolar library pool based on relative qPCR concentrations of the individual libraries.
17. Quantify the resulting library using the Qubit DS DNA HS assay. An additional ethanol precipitation step may be required as for accurate Qubit measurement the concentration should be above 1.5–2 ng/μL.
18. Sequence the library at a final concentration of 1.2 pM on a NextSeq 500 using a NextSeq 500 mid or high output V2 kit (*see* **Note 6**).

3.2 RT-qPCR

1. Thaw RNA and reverse transcriptase mix on ice. Thaw other RT II kit components at room temperature (15–25 °C). It is important to keep RNA on ice at all times to prevent degradation. Mix each solution by flicking tubes and centrifuge briefly to collect liquid from the sides and then store on ice. Do not vortex RNA.
2. Dilute RNA samples to a concentration of 100 ng/μL (*see* **Note 7**). Sensitivity can be improved by increasing the amount of input RNA.
3. Prepare RT mix by combining 2 μL of HiFlex or HiSpec buffer with 1 μL of 10× miScript Nucleics mix, 1 μL miScript Reverse Transcriptase Mix, and 4 μL of nuclease-free water (*see* **Notes 8** and **9**). If processing multiple samples, distribute 8 μL of this mixture in separate tubes and add 2 μL of RNA to each tube.
4. Incubate the reverse transcription mixture at 37 °C for 60 min followed by 5 min at 95 °C to inactivate miScript Reverse Transcriptase Mix. Place on ice.
5. Dilute the RT product 22-fold by adding 210 μL nuclease-free water (*see* **Note 10**).
6. Thaw 2× QuantiTect SYBR Green PCR Master Mix, 10× miScript Universal Primer, 10× miScript Primer Assay and template cDNA at room temperature (15–25 °C). Mix the individual solutions and place on ice.
7. Prepare the PCR mix by combining 5 μL 2× Quantitect SYBR Green PCR Master mix, 1 μL of 10× Universal Primer, 1 μL of 10× Primer assay, and 2 μL of nuclease-free water per reaction. If preparing for multiple reactions simply multiply by the number of reactions +10%. Mix by pipetting up and down and briefly spin.
8. Dispense 9 μL PCR mix in the wells of the reaction plate and add 1 μL of diluted cDNA to each reaction well. Seal reaction

wells carefully and centrifuge for 1 min at 1000 × *g* to remove bubbles.

9. Run the PCR reaction as follows: 95 °C for 15 min, (94 °C for 15 s, 55 °C for 30 s, 70 °C for 30 s) × 40 cycles, melting curve analysis.

3.3 Normalization of RT-qPCR Data

1. The normalized relative quantity for miRNA i in sample j is defined as:

$$\mathrm{NRQ}_{i,j} = 2^{\left(\mathrm{Cq}_{i,j} - \mu_j\right)},$$

with μ corresponding to either the global mean expression value or the arithmetic mean of multiple stable reference miRNAs (*see* **Note 11**), assuming 100% PCR efficiency. The qbase + software (http://www.qbaseplus.com) is particularly well suited for qPCR data analysis, including global mean or multiple reference gene normalization, and PCR efficiency correction if needed.

2. To identify stably expressed reference miRNAs: import normalized miRNA expression data into a spreadsheet like MS Excel. Calculate the standard deviation for each miRNA and select candidate miRNAs that have the lowest standard deviation, expressed in all samples and do not belong to the same miRNA family (*see* **Note 12**). Select between five and eight miRNAs as candidate reference genes. Verify in an RT-qPCR experiment that these candidate reference miRNAs are stably expressed. This means they should have low M values when using the geNorm algorithm (*see* **Note 13**).

3. In case you do not have access to miRNA-profiling data, we recommend to sequence a few representative samples followed by the procedure described above.

4. In case you do not have access to miRNA-profiling data and **step 2** is not an option, you can set up a classic geNorm pilot experiment using published candidate miRNA reference genes. Typically, eight candidate references small RNAs are evaluated in at least ten representative samples. Use the geNorm algorithm to identify the most stably expressed reference genes.

4 Notes

1. DESeq2 is available as an R package at the Bioconductor depository (www.bioconductor.org).

2. We strongly encourage using the following annotation schemes when reporting miRNA findings: the miRNA sequence itself, the miRNA name in combination with the miRBase version, or

the miRNA accession number in combination with the miRBase version.

3. To guarantee successful miRNA quantification, it is evident that the small RNA fraction is retained after RNA isolation. Several commercial kits are available that enable the extraction of total RNA including the small RNA fraction. Make use of microfluidics-based electrophoresis systems such as the Bioanalyzer or the Experion to evaluate the presence of the small RNA fraction. We strongly advise to only include RNA samples of sufficient quality. When performing miRNA quantification of total cell-free RNA present in serum or plasma samples, designated RNA isolation kits are available from different vendors. We have good experience with the miRNeasy serum/plasma kit (Qiagen).
4. When preparing a mixture for multiple samples simultaneously we advise to always prepare mixture for an additional 10%.
5. In the case of very-low input samples (e.g., serum or plasma), the number of PCR cycles can be further increased up until 16.
6. To determine optimal sequencing depth, it is possible to perform a saturation analysis in a pilot experiment by sequencing a small number of representative samples. After standard data processing the R package subSeq (available at www.bioconductor.org) can be used to determine whether enough reads were generated to detect all relevant biological information, or whether it's possible to multiplex more samples and thus work with fewer reads [12]. We typically aim for ten million reads for fresh tissue or cellular RNA, and 15 million reads for FFPE tissue or body fluid samples.
7. Concentration is dependent on the abundance of the mature miRNA target, ensure between 10 ng and 2 μg of RNA as input for the RT reaction. After RT, samples should be diluted to ensure between 25 pg and 1.5 ng of cDNA per PCR reaction.
8. In the miScript II RT kit, two buffers are included: the miScript HiSpec Buffer and the miScript HiFlex Buffer. The HiSpec Buffer is specifically formulated to facilitate the selective conversion of mature miRNAs into cDNA. The HiFlex buffer promotes the conversion of all RNA species into cDNA to enable combined study of miRNA and other RNA species like mRNA.
9. Besides dNTPs, rATP, and oligo-dT primers, the miScript Nucleics Mix contains an internal synthetic RNA control, the miRNA reverse transcription control (miRTC), that can be used to assess reverse transcription performance (i.e., absence of inhibitors).
10. To enable miRNA profiling studies of single cells and body fluids we advise to include a limited-cycle preamplification step

to increase the sensitivity of the RT-qPCR reaction. In the preamplification procedure, miRNA specific forward primers and universal reverse primers are used to amplify the cDNA template in a limited-cycle PCR, typically 12–15 cycles.

11. Baseline and threshold settings should be carefully evaluated when determining Cq-values. Typically, the baseline should be set to the cycle interval where no amplification takes place. The threshold is set, with the Υ-axis in log-scale, where all assays are in log linear phase.
12. miRNA families can be inspected in a dedicated miRBase file (ftp://mirbase.org/pub/mirbase/CURRENT/miFam.dat.gz).
13. Besides the tissue or disease type, the stability of candidate reference miRNAs also depends on the experimental conditions (e.g., treatment of the cells with siRNA or compound). We therefore advise to verify the stability of reference miRNAs when changing experimental conditions by measuring their expression on a representative selection of samples followed by geNorm analysis.

References

1. Mestdagh P, Hartmann N, Baeriswyl L et al (2014) Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. Nat Methods 11:809–815. doi:10.1038/nmeth.3014
2. Tian G, Yin X, Luo H et al (2010) Sequencing bias: comparison of different protocols of MicroRNA library construction. BMC Biotechnol 10:64. doi:10.1186/1472-6750-10-64
3. Baran-Gale J, Kurtz CL, Erdos MR et al (2015) Addressing bias in small RNA library preparation for sequencing: a new protocol recovers MicroRNAs that evade capture by current methods. Front Genet 6:1506. doi:10.1093/nar/gkr1263
4. Garmire LX, Subramaniam S (2012) Evaluation of normalization methods in mammalian microRNA-Seq data. RNA 18:1279–1288. doi:10.1261/rna.030916.111
5. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:31. doi:10.1186/s13059-014-0550-8
6. Benes V, Castoldi M (2010) Expression profiling of microRNA using real-time quantitative PCR, how to use it and what is available. Methods 50:244–249. doi:10.1016/j.ymeth.2010.01.026
7. Shi R, Chiang V (2005) Facile means for quantifying microRNA expression by real-time PCR. Biotech 39:519–525. doi:10.2144/000112010
8. Vandesompele J, De Preter K, Pattyn F et al (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 3:RESEARCH0034
9. Mestdagh P, Van Vlierberghe P, De Weer A et al (2009) A novel and universal method for microRNA RT-qPCR data normalization. Genome Biol 10:R64. doi:10.1186/gb-2009-10-6-r64
10. D'haene B, Mestdagh P, Hellemans J, Vandesompele J (2011) miRNA expression profiling: from reference genes to global mean normalization. In: Methods in molecular biology. Humana Press, Totowa, NJ, pp 261–272
11. Van Peer G, Lefever S, Anckaert J et al (2014) miRBase tracker: keeping track of microRNA annotation changes. Database 2014:bau080. doi:10.1093/database/bau080
12. Robinson DG, Storey JD (2014) subSeq: determining appropriate sequencing depth through efficient read subsampling. Bioinformatics 30:3424–3426. doi:10.1093/bioinformatics/btu552

Chapter 14

Using FirePlex™ Particle Technology for Multiplex MicroRNA Profiling Without RNA Purification

Michael R. Tackett and Izzuddin Diwan

Abstract

Accuracy of miRNA profiling is enhanced when sample processing can be kept to a minimum, avoiding steps such as RNA purification that can introduce bias and inaccuracies. Here we describe a novel multiplex circulating miRNA assay that enables the profiling of up to 65 miRNAs of choice in the same well directly from plasma (including heparin plasma) or serum, with no need for RNA purification. The main component of the assay is FirePlex™ hydrogel particles, which enable the multiplex capture of miRNAs with picomolar sensitivity and high specificity. Results are obtained using conventional flow cytometry and easy to use software, which allows fast analysis and interpretation of the experimental data. This chapter provides methods to profile miRNAs with PCR sensitivity from as little as 10 μL of crude biofluid sample, or from less than 100 pg of purified RNA.

Key words microRNA, Multiplex, High-throughput, Polymerase chain reaction, Hybridization

1 Introduction

MicroRNAs (miRNAs) are a class of 18–25 nucleotide noncoding RNAs that regulate protein expression upon binding to appropriate mRNAs by blocking their translation and/or mediating their degradation [1, 2]. A single miRNA frequently targets dozens of genes [3] and the majority of human genes are regulated by miRNAs [4]. They have been shown to be involved in many biological processes, including development, proliferation, signal transduction, differentiation, and apoptosis [5]. As such, researchers have investigated their potential as biomarkers in many diseases, including cancers [6], neurodegenerative diseases [7], cardiovascular disease [8], and liver disease [9].

Multiple factors make miRNAs well suited to serve as biomarkers in "liquid biopsies," whereby miRNA profiling in a biofluid reflects disease state within the body. First, miRNAs are released by cells in fairly stable forms, either bound to proteins such as Argonaute [10] or packaged in various extracellular vesicles [11], and

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_14,

remain stable even after years of sample storage [12]. Second, miRNAs have been consistently profiled in a large number of biofluids, including plasma, sera, urine, and saliva [13]. As a result, circulating miRNAs have been studied for many purposes, including disease diagnosis, disease stratification, and as companion diagnostics.

However, the relatively low abundance of miRNAs in biofluids, and the diversity of potential sample types and sample collection methodologies present challenges for using circulating miRNAs as biomarkers [14]. Additionally, methods for sample QC are often inadequate. Most existing technologies require relatively large volumes of initial sample (>200 μL) from which miRNAs are isolated. The resulting miRNA is often of low quantity, necessitating a pre-amplification step which can introduce bias. Finally, many existing profiling technologies lack the ability to affordably profile large numbers of miRNAs in a large number of samples. There is therefore a need for miRNA profiling techniques that are sensitive, specific, reproducible, high-throughput, and customizable, all while profiling directly from the sample of interest.

In this chapter we describe a method that addresses these needs through the application of encoded hydrogel microparticles [15–17]. Up to 65 miRNAs are simultaneously profiled from as little as 10 μL of biofluid, in a single well of a 96-well filter plate. With the FirePlex™ technology, barcoded particles containing probes complimentary to the miRNA of interest are generated using poly(ethylene glycol) hydrogel. This substrate is porous (allowing higher sensitivity due to the use of its full three-dimensional structure) and non-fowling (allowing its use directly in biofluids).

As illustrated in Fig. 1, samples are first digested in lysis buffer, then hybridized directly to the mixture of hydrogel particles, where miRNAs present in solution specifically hybridize to a complimentary probe. After rinses, adapters are ligated onto both ends of the hybridized miRNA molecule, and the fusion molecule is then eluted from the particle for PCR amplification. Importantly, this first hybridization step also functions as a means of isolation, thereby allowing PCR inhibitors such as heparin to be rinsed

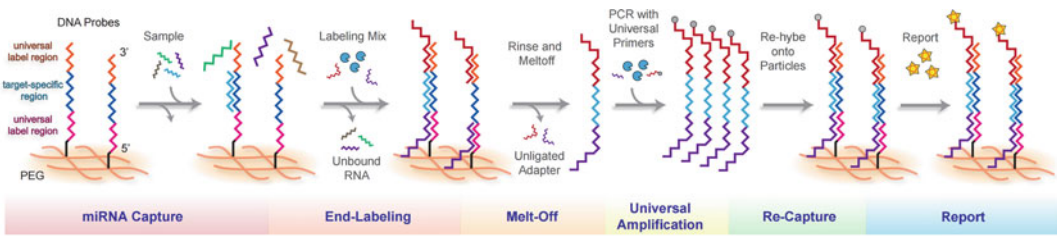


Fig. 1 Schematic overview of the molecular biology of the assay

away, preventing them from reducing the sensitivity of the assay [18]. The PCR uses a biotinylated primer to tag the amplicon before the amplicon is re-hybridized back onto the particles. Finally, the particles are incubated with a fluorescent reporter and scanned on a standard flow cytometer. Included software is used to deconvolute the codes and determine the fluorescence intensity of a given microparticle. Up to 30 particles of each type are present in a given well, allowing for improved statistical accuracy.

2 Materials

2.1 Reagents

1. RNAse/DNAse-free water.
2. FirePlex™ miRNA Assay Core Reagent Kit (Abcam).
   (a) 1× Hybridization buffer.
   (b) 10× Rinse buffer A.
   (c) Labeling diluent.
   (d) 2× Rinse buffer B.
   (e) Run buffer.
   (f) Filter plate.
   (g) Catch plate.
   (h) Protease mix.
   (i) Digest buffer.
   (j) Labeling buffer.
   (k) Labeling enzyme.
   (l) PCR buffer.
   (m) dNTP mix.
   (n) Primer mix.
   (o) PCR enzyme.
   (p) 5× Reporter solution.
   (q) Control RNA.
3. FirePlex™ miRNA Assay Immunology Panel (Abcam).

2.2 Equipment

1. Vacuum manifold.
2. Heated orbital shaker (*see* **Note 1**).
3. Thermocycler.
4. Cytometer (*see* **Note 1**).
5. 25 mL reservoirs.
6. PCR strips or plate.
7. Multichannel pipettes.

2.3 Analysis Software

1. FirePlex™ Analysis Workbench (http://www.abcam.com/FireflyAnalysisSoftware).

3 Methods

Carry out all procedures at room temperature unless otherwise indicated.

3.1 Reagent Preparation

Dilute 10× rinse buffer A by mixing the entire supplied volume (30 mL) with 270 mL RNase-free water in a clean container. Dilute 2× Rinse buffer B by mixing the entire supplied volume (33 mL) with 33 mL RNase-free water in a clean container. Excess 1× rinse buffers can be stored at room temperature.

3.2 Sample Preparation

Biofluid samples can be profiled without processing, though brief centrifugation to clarify the sample can minimize variation (*see* **Note 2**). For best results, samples should be stored at −80 °C and the number of freeze/thaw cycles limited. This method also successfully profiles miRNAs in isolated total RNA. The percentage of miRNA in a total RNA sample, and the relative abundance of the different miRNAs within a sample, may vary between isolation methods.

3.3 Lysis

1. Prepare lysis buffer according to manufacturer's instructions: prepare 44 μL per sample to be run by mixing 40 μL digest buffer with 4 μL protease mix, and adjusting to the number of samples.
2. Optimal sample digestion occurs when 40 μL of the prepared lysis buffer is mixed with sample in a well of a sterile user-supplied PCR plate or 8-well PCR tube strip, with a total final volume of 80 μL. To supplement the volume up to the recommended 80 μL, use RNase-free water. Table 1 illustrates recommended volumes for various biofluids, but the optimal volume for your sample may vary based on many factors. Use RNase-free water as sample input for negative control wells.
3. Carefully seal the tubes or plate and incubate the samples for 45 min at 60 °C while shaking (*see* **Note 3**).

Table 1
Recommended volumes for various biofluids

Sample type	Sample volume (μL)	Water volume (μL)	Lysis buffer (μL)
Human amniotic fluid, human bile, mouse sera, mouse plasma, breast milk	10	30	40
Human plasma, human sera	20	20	40
Saliva, urine	40	0	40

4. Remove samples from shaker and store in freezer until needed (*see* **Note 4**).
5. Adjust the temperature of the shaker to 37 °C.

3.4 Hybridization

1. Check that a heated shaker is at 37 °C (*see* **Note 3**).
2. Peel backing off the plate seal and apply over the filter plate (not the filter plate lid). Cut off the seal from the included filter plate to reveal one well for each sample and one well for each control (*see* **Note 5**).
3. Invert FirePlex particles end-over-end and vortex before adding 35 µL to each well of filter plate, keeping particles mixed while distributing. Mixing is vital to ensure that each well receives an equal number of particles. Close and re-invert FirePlex particles tube every five wells (*see* **Note 6**).
4. Apply vacuum to the filter plate to remove storage buffer and blot the underside of the plate dry with a Kimwipe™. Excess buffer under the filter plate wells may result in assay failure, so blot thoroughly (*see* **Note 7**).
5. Add 25 µL hybridization buffer to each well of the filter plate. Hybridization buffer is viscous; take care during pipetting to ensure each well receives an equal volume.
6. Transfer 25 µL digested sample to each well of the filter plate. Alternatively, transfer 25 µL total RNA if you have isolated RNA to run. As a positive control, dilute 1 µL of the Control RNA included with the kit into 24 µL RNase-free water and load that (*see* **Note 8**).
7. Cover with lid and incubate the samples for 60 min at 37 °C while shaking (*see* **Note 3**).

3.5 Labeling

1. Remove filter plate from shaker and adjust the temperature of the shaker to room temperature. Alternatively, a second shaker may be used.
2. For a single well, prepare 1× labeling mix by combining 78.4 µL labeling diluent, 1.6 µL labeling buffer, and 0.4 µL labeling enzyme. Vortex to mix.
3. Rinse wells by applying 165 µL 1× rinse A on top of the liquid in each well followed by application of vacuum.
4. Rinse wells a second time by applying 165 µL 1× rinse A to each well followed by application of vacuum. Blot the underside of the plate dry.
5. Add 75 µL 1× labeling mix prepared above to each well.
6. Cover filter plate with lid and incubate the samples for 60 min at room temperature while shaking.

3.6 PCR

1. Adjust the temperature of the shaker to 55 °C.
2. Thaw −20 °C PCR reagents and store on ice.
3. Rinse wells by applying 165 μL 1× rinse B on top of liquid in each well followed by application of vacuum.
4. Rinse wells a second time by applying 165 μL 1× rinse B directly to particles in each well followed by application of vacuum.
5. Rinse wells once by applying 165 μL 1× rinse A to the wells followed by application of vacuum. Be sure to use 1× rinse A for this step. After final rinse, blot the underside of the plate to remove excess liquid.
6. Add 110 μL RNase-free water to each well.
7. Cover filter plate with lid and incubate the samples for 30 min at 55 °C while shaking.
8. Insert the catch plate into the vacuum manifold and place the filter plate on the vacuum manifold, aligning carefully (Fig. 2). Then apply suction, catching eluent in the filter plate. Note the orientation of the catch plate so that the proper samples get transferred to PCR (*see* **Note 9**).
9. Remove the filter plate from the vacuum manifold, blot the bottom if necessary, and add 175 μL 1× rinse A to each well. Cover the filter plate with its lid and store at 4 °C until it is needed after the PCR.
10. For a single well of PCR, prepare PCR master mix by combining 19.8 μL PCR buffer, 2.4 μL primer mix, 1.2 μL dNTP mix, 0.6 μL PCR enzyme in that order and mixing well. Store on ice until ready for use.

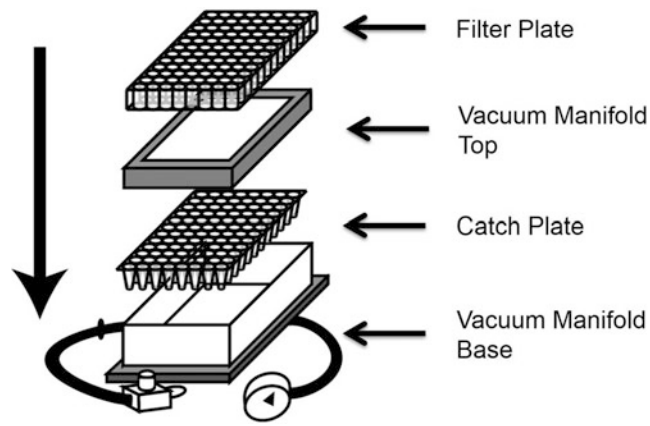


Fig. 2 Schematic overview of how to set up filter plates to receive eluent

Table 2
Thermal cycling procedure

Cycle	Temperature and time
1 cycle	93 °C for 15 s
32 cycles	93 °C for 5 s 57 °C for 30 s 68 °C for 60 s
1 cycle	68 °C for 5 min
1 cycle	94 °C for 4 min
1 cycle	4 °C forever

11. Using a fresh user-supplied PCR plate, first mix the eluent by pipetting up and down and then transfer 30 μL of the eluent from **step 8** from the catch plate to the fresh PCR plate.
12. Add 20 μL PCR master mix to each well of the user-supplied PCR plate containing 30 μL of eluent in **step 11**, mixing well by pipetting up and down. Remember to change tips between pipetting different samples.
13. Transfer reaction mixtures to a thermal cycler.
14. Thermal cycle using the procedure listed in Table 2.

3.7 Capture

1. Adjust the temperature of the shaker to 37 °C. While the same shaker used in previous steps may be reused, to limit PCR contamination, a separate, post-PCR shaker is recommended for this and future steps.
2. Apply vacuum to the filter plate to remove the 1× rinse A that kept the FirePlex Particles stable during storage and blot the underside dry.
3. Add 60 μL hybridization buffer to each well of the filter plate, then transfer 20 μL of the PCR product from the PCR plate to the filter plate. Care should be taken to place the PCR product in the well from which its corresponding eluent was taken (*see* **Note 8**).
4. Cover with lid and incubate the samples for 30 min at 37 °C while shaking.

3.8 Report

1. Adjust temperature of heated shaker to room temperature.
2. For one well, prepare 1× reporter mix by combining 64 μL RNase-free water and 16 μL 5× reporter. Vortex to mix.
3. Remove filter plate from shaker.

4. Rinse wells by applying 165 μL 1× rinse B on top of liquid in each well followed by application of vacuum. Rinse wells a second time with by applying an additional 165 μL 1× rinse B to each well followed by application of vacuum.
5. Rinse wells by applying 165 μL 1× rinse A to each well followed by application of vacuum. After rinse, blot the underside of the plate dry.
6. Add 75 μL 1× reporter mix prepared above to each well.
7. Cover filter plate with lid and incubate the samples for 15 min at room temperature while shaking.

3.9 Scan

1. Rinse wells by applying 165 μL 1× rinse A to the top of each well followed by application of vacuum.
2. Rinse wells a second time by applying 165 μL 1× rinse A to the top of each well followed by application of vacuum. After rinse, blot the underside of the plate dry.
3. Add 175 μL run buffer to each well (do not mix).
4. Ensure that the wells aren't leaking by setting the filter plate on a dry surface to see if there is liquid transfer after 30 s. If leakage occurred simply re-blot the underside and bring the volume of run buffer up to 175 μL in the leaky wells and reassess.
5. Scan on an approved flow cytometer (*see* **Note 10**).

3.10 Analysis

1. Launch the FirePlex Analysis Workbench and load the .fcs file (s) associated with your run.
2. Provide the panel barcode or plex file information to the software to indicate which mixture of miRNAs was profiled.
3. Within the software, create an experiment containing your samples of interest.
4. Heat maps are generated to present the expression of all targets in all wells (Fig. 3) and data can be exported in .csv format and analyzed to compare between groups (Fig. 4).

4 Notes

1. Suitable shakers have an orbital diameter of 3 mm and shake at 750 rpm. For shakers with a different orbital diameter, adjust the rpm according to the following formula: rpm = $(1{,}687{,}500/\text{orbital diameter of shaker in mm})^{½}$. Information about criteria for cytometer compatibility may be found on the Abcam webpage.
2. It isn't necessary to do so, but for improved sample reproducibility samples may be first centrifuged for 10 min at 2000 × *g* to remove all cellular debris.

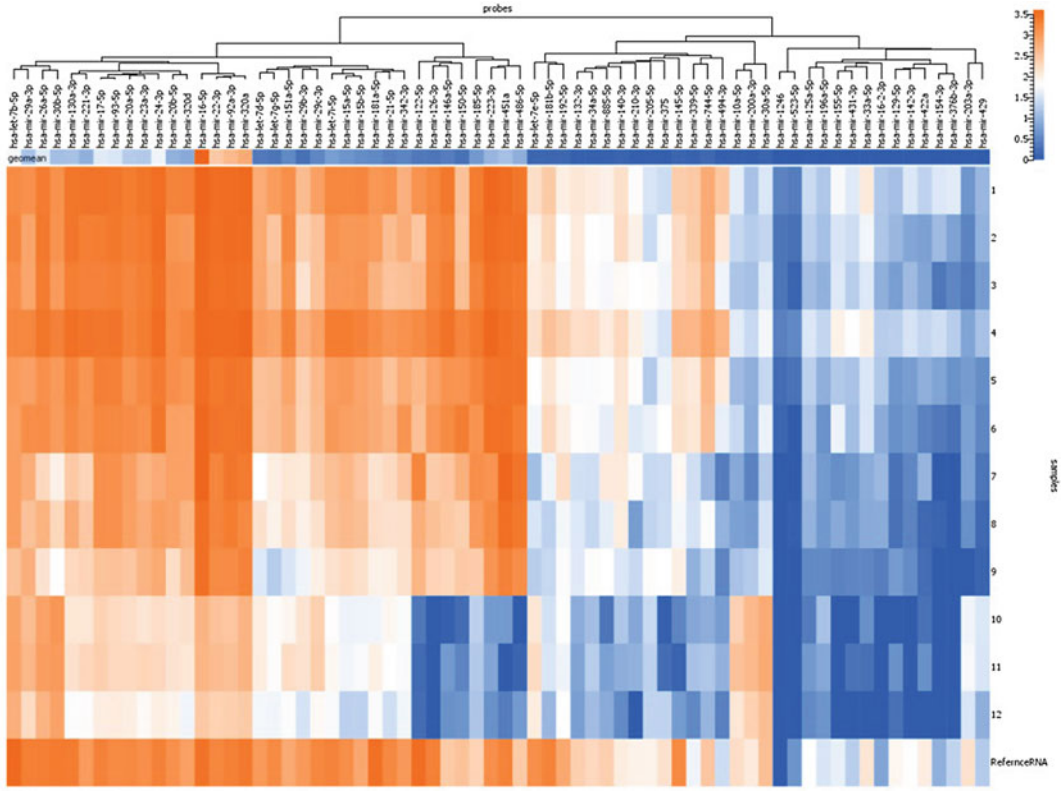


Fig. 3 Heat map showing relative abundance of circulating miRNAs in a number of biological fluids, shown in log10 of mean fluorescence intensity. Blood was drawn from three individuals into each of three preparation tubes. Urine exosomes were also profiled. Samples 1–3: 20 μL K2 EDTA plasma. Samples 4–6: 20 μL Na heparin plasma. Samples 7–9: 20 μL serum. Samples 10–12: urine exosomes, Reference RNA: 1 ng Control RNA

3. We recommend you use a second, external thermometer to independently verify that the heated shakers, when set both reach and maintain those temperatures. It is important that steps be performed at the recommended temperatures.
4. Unused digested sample can be stored for at least 2 weeks at −80 °C.
5. Do not reapply seal at any point (when covering plate, do so only with the supplied lid). If the plate seal is reapplied to the plate it may result in leakage during subsequent plate shaking.
6. If distributing particles with a multichannel pipette, add 4 mL 1× rinse A to a clean reagent reservoir (not supplied). Add 4 mL FirePlex particles to the reservoir, and mix by rocking ten times. Use an eight-well multichannel pipette to transfer 70 μL of the particle mix, pipette up and down once between each transfer for columns 1–10. Then remove four tips from the multichannel pipette, tilt reservoir, and transfer the remaining

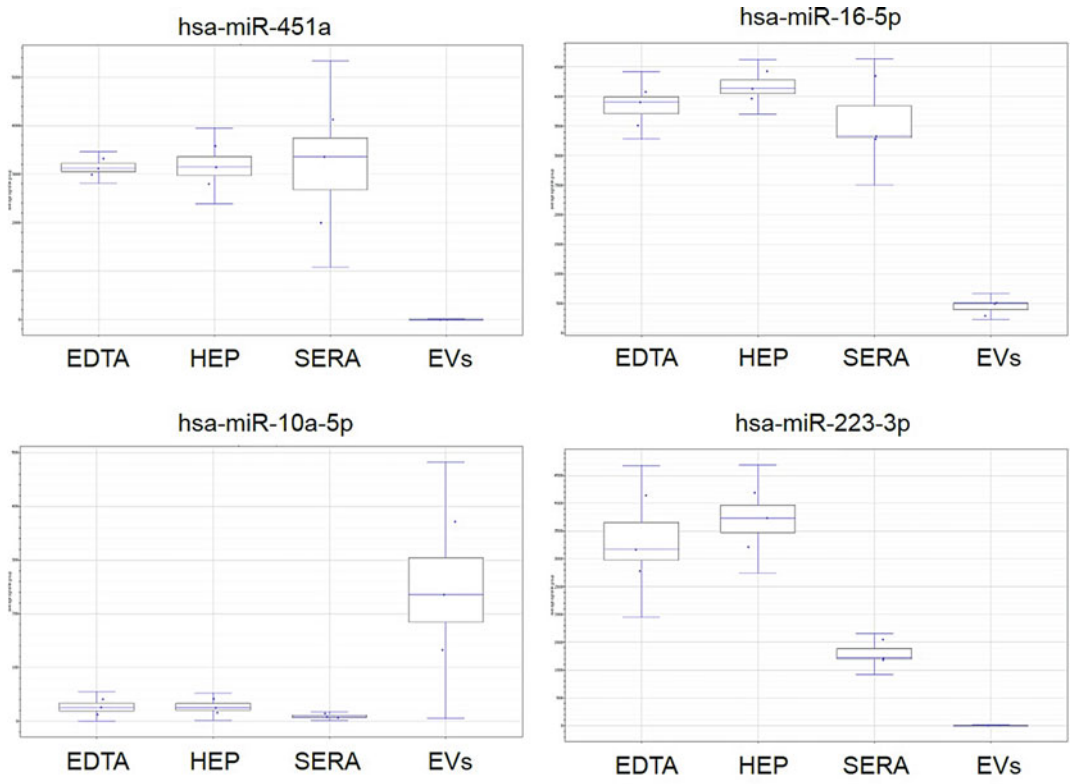


Fig. 4 Expression of four select miRNAs in different sample types. EDTA: 20 μL K2 EDTA plasma. HEP: 20 μL Na heparin plasma. SERA: 20 μL serum. EVs: urine exosomes

particle mix four wells at a time, pipetting up and down twice between transfers.

7. When applying vacuum to samples in the filter plate, turn off the vacuum as soon as the liquid is gone from each well to prevent over-drying.
8. It is important to maintain the particles, sample and hybridization buffer volumes listed herein; adjusting these volumes will negatively impact the assay. To modify input amounts you can, where possible, adjust the concentration of samples by diluting them or resuspending isolated RNA in smaller volumes.
9. Excess eluent and PCR amplicon may be stored at −20 °C for weeks without degradation and reused at a later time. This is recommended in case mistakes are made during PCR sample prep or rehybridization. Fresh particles may be used and the process continued.
10. Do not reuse filter plates. Once a filter plate contains PCR product, it should not be used to run the assay again due to the risk of cross-contamination. Use a fresh filter plate for subsequent runs if only a portion of the reagents were used.

Acknowledgments

FirePlex™ is a trade mark of Abcam PLC.

References

1. Ambros V (2003) MicroRNA pathways in flies and worms: growth, death, fat, stress and timing. Cell 113:673–676
2. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281–297
3. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120:15–20
4. Friedman RC, Farh KK, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 19:92–105
5. Huang Y, Shen XJ, Zou Q, Wang SP, Tang SM, Zhang GZ (2011) Biological functions of microRNAs: a review. J Physiol Biochem 67:129–139
6. Calin GA, Croce CM (2006) MicroRNA signatures in human cancers. Nat Rev Cancer 6:857–866
7. Grasso M, Piscopo P, Confaloni A, Denti M (2014) Circulating miRNAs as biomarkers for neurodegenerative disorders. Molecules 19:6891–6910
8. Quiat D, Olson E (2013) MicroRNAs in cardiovascular disease: from pathogenesis to prevention and treatment. J Clin Invest 123 (1):11–18
9. Szabo G, Bala S (2013) MicroRNAs in liver disease. Nat Rev Gastroenterol Hepatol 10 (9):542–552
10. Li L, Zhu D, Huang L, Zhang J, Bian Z, Chen X, Liu Y, Zhang C, Zen K (2012) Argonaute 2 complexes selectively protect the circulating microRNAs in cell-secreted microvesicles. PLoS One 7(10):e46957
11. Raposo G, Stoorvogel W (2013) Extracellular vesicles: exosomes, microvesicles, and friends. J Cell Biol 200(4):373–383
12. Balzano F, Deiana M et al (2015) MiRNA stability in frozen plasma samples. Molecules 20(10):19030–19040
13. Weber J, Baxter D, Zhang S, Huang D, Huang K, Lee M, Galas D, Wang K (2010) The microRNA spectrum in 12 body fluids. Clin Chem 56(11):1733–1741
14. Moldovan L, Batte K, Trgovcich J, Wisler J, Marsh C, Piper M (2014) Methodological challenges in utilizing miRNAs as circulating biomarkers. J Cell Mol Med 18(3):371–390
15. Pregibon D, Doyle P (2009) Optimization of encoded hydrogel particles for nucleic acid quantification. Anal Chem 81:4873–4881
16. Chapin S, Pregibon D, Doyle P (2009) High-throughput flow alignment of barcoded hydrogel microparticles. Lab Chip 9:3100–3109
17. Chapin S, Pregibon D, Doyle P (2011) Rapid microRNA profiling on encoded gel microparticles. Angew Chem Int Ed 50:2289–2293
18. Boeckel J, Thome C et al (2013) Heparin selectively affects the quantification of microRNAs in human blood samples. Clin Chem 59 (7):1125–1127

Chapter 15

Multiplex Real-Time PCR Using Encoded Microparticles for MicroRNA Profiling

Seungwon Jung and Sang Kyung Kim

Abstract

Multiplex quantitative real-time PCR (qPCR), which measures multiple DNAs in a given sample, has drawn unprecedented attention as a means of verifying the rapidly increasing genetic targets in a single phenotype. We report the detailed procedure of a readily extensible qPCR for multiple microRNA (miRNA) expression analysis using microparticles of primer-immobilized networks as discrete reactors. Individual particles are identified by two-dimensional codes engraved into the particles. It allows high-fidelity signal analysis in the multiplex real-time PCR. During the course of PCR, the amplicons accumulate in the volume of the particles with amplification efficiency over 95%. Tens of miRNAs can be quantitatively profiled in a single PCR reaction of this method.

Key words MicroRNA, Hydrogel, Real-time PCR, Multiplex, Encoded particle

1 Introduction

The quantitative profile of miRNAs has become prevalent as miRNAs have recently been recognized as novel diagnostic parameters for significant diseases such as inherited diseases, cancers, heart diseases, infection, alcoholism, and obesity [1–5]. qPCR is considered to be the gold standard in miRNA expression analysis because it is a well-characterized methodology with a wide dynamic range and low limit of detection compared to northern blot or microarray [6]. Until recently, however, the number of multiplicity in common qPCR has been limited to a maximum of six due to the restricted number of color channels [7]. Here, we present a multiplex qPCR for miRNA profiling equipped with novel particles, encoded primer-immobilized networks (PIN) [8, 9]. PIN, composed of polyethylene glycol (PEG), is highly porous and hydrophilic so that the amplification reaction in the particles is as efficient as in an aqueous medium. Each encoded PIN of a specific primer is

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_15,

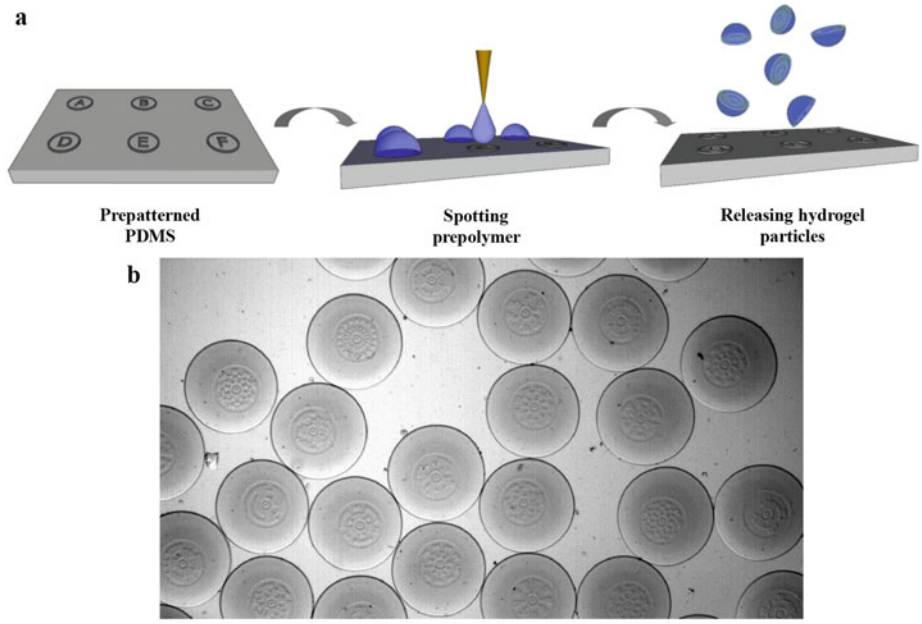


Fig. 1 The production of encoded microparticle for qPCR. (**a**) Encoded particles were produced by spotting the pre-polymer solution onto the pre-patterned PDMS surface followed by curing and releasing processes. (**b**) The code patterns were replicated from the PDMS surface to the particles

identified by an engraved pattern, which has a coding capacity of much greater than the number of human miRNAs (Fig. 1). Thus, the addition of a relevant encoded PIN readily expands the target of analysis to that allowed by the available space (Fig. 2).

2 Materials

Prepare all solutions using DNAse- and RNAse-free water and bioanalytical grade reagents. Diligently follow all waste disposal regulations when disposing waste materials.

2.1 Polyethylene Glycol Pre-polymer

1. Pre-polymer buffer: 3× TE buffer containing 0.15% Tween-20 (3× TET).
2. Polyethylene glycol diacrylate ($m_n = 700$, PEGDA): Store at 4 °C.
3. Polyethylene glycol ($m_n = 600$, PEG) (*see* **Note 1**).
4. Photoinitiator: 2-Hydroxy-2-methyl-1-phenyl-propan-1-one.
5. Acrydite primer: 200 μM in water. Modify 5′ end of specific forward primer with acrydite group.
6. Filter syringe.

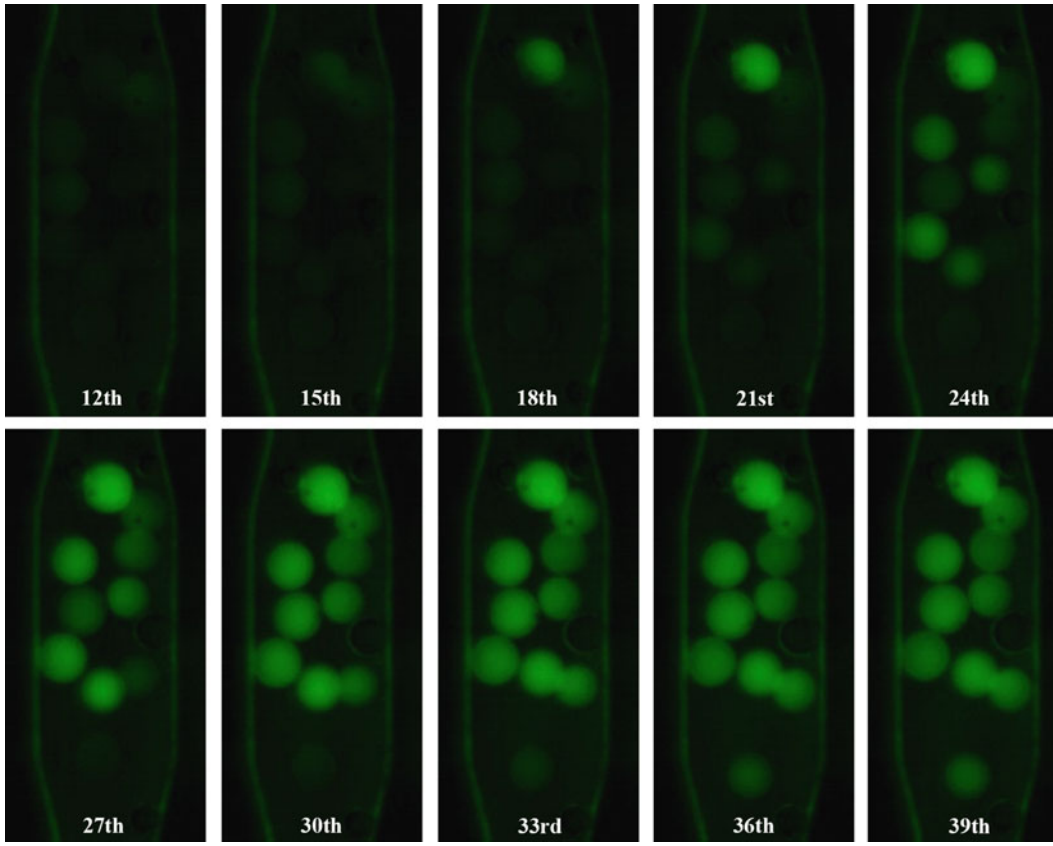


Fig. 2 Image sequences of qPCR with ten different particles. The cDNA samples reverse-transcribed from the K562 cell line extracellular vesicles were inserted into the PCR chamber, ten different particles were placed. Each particle showed independent amplification according to the amount of the target cDNAs

2.2 Encoding Mold

1. Glass mask: the chromium patterns on the glass (*see* **Note 2**).
2. Silicon wafer: 4-in. wafer with single-side chemical mechanical polishing. Its crystalline orientation does not matter.
3. Piranha solution: 99% sulfuric acid/hydrogen peroxide (4:1): Mix 200 mL sulfuric acid and 50 mL hydrogen peroxide in the Teflon bowl (*see* **Note 3**).
4. Buffered Oxide Etchant: 40% NH_4F/49% HF (6:1) (BOE-Solution).
5. N_2-Gun.
6. Spin Coater.
7. SU-8 2005 (Microchem, Westborough, MA, USA).
8. SU-8 developer (Microchem, Westborough, MA, USA).
9. Isopropyl alcohol (IPA).
10. Teflon bowl.
11. Stainless steel bowl.

12. Plastic Petri dish with 15 cm in diameter.
13. Trichloro(1H, 1H, 2H, 2H-perfluorooctyl)silane.
14. Scotch tape.
15. Polydimethylsiloxane (PDMS): silicone elastomer base and silicone elastomer curing agent.
16. 80 °C oven.
17. 95 °C/120 °C hotplate.

2.3 Particle Production

1. Microspotting system (*see* **Note 4**).
2. UV chamber.
3. Rinsing buffer: 1× TET produced by dilution of 3× TET.
4. Mini centrifuge.
5. Vortexer.

2.4 Reverse Transcription and qPCR

1. RnaseZAP: cleaning spray for handling RNAs (Sigma-Aldrich).
2. RT kit: QuantiMir RT Kit (System Biociences, Palo Alto, CA, USA) (*see* **Note 5**). Store at −20 °C.
3. qPCR mastermix: NBS SYBR Green REAL-TIME PCR KIT (Nanobiosys, Seoul, South Korea) or SSOFast™ EvaGreen® Supermix (Bio-Rad, Hercules, CA, USA). Store at −20 °C.
4. Universal reverse primer: 10 μM in water (System Biosciences, Palo Alto, CA, USA). Store at −20 °C.
5. PCR equipment and chamber: UltraFast LabChip Real-time PCR G2-3 system (Nanobiosys, Seoul, South Korea).
6. Standard DNA: synthesized amplicon with known concentration corresponding to primers.

3 Methods

3.1 Encoding Mold

1. Heat the piranha solution up to 120 °C and dip 4-in. silicon wafer into the solution for 10 min in order to clean the surface of the wafer. Rinse the wafer thoroughly in flowing distilled water for 6 min.
2. Dip the wafer into the BOE solution for 10 s. Rinse the wafer thoroughly in flowing distilled water for 6 min. Dry it carefully using N_2 gun.
3. Place the wafer on the hotplate pre-heated to 120 °C for 5 min in order to dehydrate it perfectly.
4. After cooling the wafer in room condition, place it on the spin coater and turn vacuum on. Pour SU-8 2005 on the wafer. Spin it under the condition of 13 × *g* for 10 s and 900 × *g* for 40 s in order to obtain 5 μm thickness (*see* **Note 6**).

5. Place the wafer on the 95 °C hotplate for 2 min in order to evaporate the solvent. Cool the wafer slowly (*see* **Note 7**).
6. Place the glass mask on the wafer.
7. Expose the masked PR-coated wafer to UV on with the energy of 120 mJ/cm^2.
8. Place the wafer on the 95 °C hotplate for 3 min again. Cool the wafer slowly (*see* **Note 7**).
9. Dip the wafer into stainless steel bowl filled with the SU-8-developer with mild agitation for 1 min. Rinse it with isopropyl alcohol thoroughly and dry it with N_2 gun.
10. Bake the wafer at 150 °C for 5 min to enhance its mechanical property.
11. Place the wafer in vacuum chamber with 1 μL drop of Trichloro(1H, 1H, 2H, 2H-perfluorooctyl)silane around the wafer for coating on the surface of SU-8 and silicon with hydrophobic self-assembled monolayer (SAM) (*see* **Note 8**). Attach the wafer at the bottom of the Petri dish with scotch tape.
12. Mix 50 g silicone elastomer base and 5 g silicone elastomer curing agent into paper cup till being intransparent due to bubbles. Keep it in vacuum chamber until the bubbles are removed (approx. 1 h). Pour the mixture into the wafer-loaded Petri dish and remove the bubbles in vacuum chamber after planarization. Put it in the 80 °C oven for 1 h (*see* **Note 9**).
13. Detach the cured silicone from the wafer and cut into pieces according to the pattern. Place each silicone piece on slide glass.

3.2 Primer

1. Design each forward primer for miRNAs according to the sequence of mature miRNAs (*see* **Note 10**).
2. Validate the primer performance with the qPCR in solution phase (*see* **Note 11**).

3.3 Particle Production

1. Mix 20 μL of PEG700DA, 40 μL of PEG600, 35 μL 3× TET buffer, and 5 μL photoinitiator in a 200 μL tube (*see* **Note 12**). Vortex for 10 s to mix thoroughly and centrifuge at 2000 × *g* for 1 min. Remove impurities in the pre-polymer solution using filter syringe. Complete the PCR pre-polymer solution by mixing 90 μL pre-polymer solution and 10 μL 200 μM acrydite primer. Repeat vortexing and centrifugation. Wrap in aluminum foil before use.
2. Place the patterned silicone-laid slide glass on the microspotting system. Spot the pre-polymer solution on the patterns of the silicone surface with about 25 nL per each spot after

aligning the position at three points of silicone surface (*see* **Note 13**) (Fig. 1a).

3. Move the silicone-laid glass slide with spotted pre-polymer droplets into the UV chamber and cure it with the power of 43 mJ. Pour 1× TET on the surface and collect the photo-crosslinked particles by pipetting on the surface (*see* **Note 14**).
4. Spin down the particles at 2000 × *g* for 5 min, remove supernatant, and leave 100 μL. Add 900 μL of 1× TET buffer and vortex for 10 s. Repeat five times. Finally, remove supernatant and leave 100 μL (*see* **Note 15**).

3.4 Reverse Transcription

1. Clean the experimental table and hand gloves with RnaseZAP.
2. Extract total RNA from the samples (*see* **Note 16**).
3. Follow manufacturer's instruction (QuantiMir RT Kit).

3.5 qPCR with Multiple Particles

1. Prepare the PCR solution: 8 μL 2× qPCR Mastermix, 1 μL cDNA solution, 1.6 μL 10 μM universal reverse primer, 5.4 μL water for 16 μL in total (*see* **Note 17**).
2. Aspire the mixture solution of target-specific particles using the 1 mL pipette and then insert these particles into the planar-type PCR chamber (*see* **Note 18**).
3. Fill the chamber with PCR solution (*see* **Note 19**).
4. Confirm the information of particles according to the codes under the microscope.
5. Load the PCR chamber and determine the region of interest for measuring the fluorescence change at each particle.
6. Set the adequate condition for qPCR of multiple cDNAs reverse-transcribed from miRNAs (*see* **Note 20**).
7. Run the process after loading PCR chamber. Obtain the fluorescence data from the particles and draw the graph to represent the change in fluorescence intensity according to the cycle number (Fig. 2).
8. In order to evaluate the particle-based qPCR, run the PCR process without template and tenfold serial-diluted standard DNAs. Plot the threshold cycle versus the dilution factor (standard curve) and fit the data to a straight line (*see* **Note 21**).

4 Notes

1. PEG can be used as a porogen to make the pores in the particle. According to porogen molecular weight and percentage, the pores can be tuned [10]. In our lab, we decided 40% PEG600 to be an optimal porogen composition for qPCR.

2. We tested diverse pattern sizes for encoded particles. The 20 μm feature size is the minimum for replica pattern to the particles (Fig. 1b).
3. Wear protection clothes, gloves, and mask. The reaction should be carried out in allowed area such as acid fume hood. When sulfuric acid and hydrogen peroxide are reacted, pour carefully sulfuric acid first and hydrogen peroxide later because it is an exothermal process.
4. The microspotting equipment for nL-scale droplet formation is necessary. Since the pre-polymer solution we used is viscous, syringe-based spotter (solenoid) is more suitable than piezo type. We purchased the syringe-based microspotting system named Arrayer 2000 (Advanced Technology Inc., Incheon, South Korea).
5. We used the polyA-tailing method for reverse transcription of miRNAs. This method reverse-transcribes every miRNA and every RNA as well. If you want to see the profile of some miRNAs more accurately, use the stem-loop RT kit to reverse-transcribe certain miRNAs according to inserted RT primers.
6. When pouring the SU-8 photoresist, be careful to avoid the generation of air bubbles in the photoresist. The air bubbles deteriorate the coating uniformity and lead to defects in the final pattern. The spin condition and following protocols can be changed in order to obtain the target thickness.
7. Quick temperature change may induce thermal stress on the SU-8, leading to the detachment of SU-8 from the wafer. Whenever changing the temperature of the wafer, the temperature should be changed slowly. In our lab, we heated up the temperature of the hotplate after loading the wafer and turned it off for cooling.
8. SAM solution is very harmful. Wear gloves and mask when handling. The vacuum chamber for SAM coating should be separated inside a second vacuum chamber because SAM will coat the chamber on all surfaces.
9. Keeping the Petri dish in parallel with the bottom is important during the curing process. Since PDMS before curing is fluidic, final PDMS thickness can vary depending on the position if the Petri dish is not parallel with bottom.
10. The sequence of the primer can be minutely modulated for getting the melting temperature similar and increasing specificity. Furthermore, special primers like locked nucleic acid (LNA) can be used for a more specific assay.
11. Before using the qPCR in particle phase, check the amplification performance of designed primers in solution phase. Carry

out serially diluted samples and no-template control (NTC), respectively. From serial dilution, we can calculate the PCR efficiency according to $\varepsilon_{PCR} = -1 + 10^{(-1/slope)}$, where the slope is produced by a qPCR standard curve (the threshold cycle versus the dilution factor) [11]. PCR efficiency should be larger than 85% and Ct value for NTC should be later than 30. If not, redesign the primers for enhancing the performance.

12. The phase of PEG600 varies according to the surrounding temperature because its melting temperature is 20–25 °C. It is solid phase in the winter and gel phase in the summer. In order to confirm its volume, it should be melted by hands or heater after filling the tube with it using a spatula. In addition, since it is still viscous even after fully melting, pipetting should be very slow to aspirate the required amount. PEG700DA is a strong gel phase because of the storage in 4 °C. Like PEG600, it can be used after melting.
13. The humidity of the spotting chamber is important for quality control of the spotted particles. If the humidity is low, the pre-polymer solution spotted earlier is evaporated and its composition ratio is also changed. On the contrary, high humidity can change the composition as pre-polymer absorbs its surrounding water either. Thus, the humidity in chamber during spotting should be maintained between 50 and 60% RH for uniform properties of the particles.
14. UV treatment causes thymine-thymine cyclobutane dimer between neighboring thymines [12]. In order to see the effect of UV dose, we carried out the particle curing with different UV doses. As a result, the PCR performance was not changed according to the UV dose that we used (21–260 mJ).
15. Rinsing protocol for removal of porogen and unbound primers is important for reliability of the results because remaining primers can distort the fluorescence signal and hinder the local accumulation of fluorescence in the particle [8].
16. Total RNA can be extracted from the various samples such as cell line, tissue, and other biopsy. Depending on the purpose, select the sample and prepare the total RNA. In order to reduce the nonspecific reaction, the small RNA purification can be used. Total RNA should be reversely transcribed into the cDNA right after extraction in order to avoid the degradation. If not possible, store it in −80 °C till the use.
17. The PCR solution should be kept on ice before the reaction and protected from light.
18. When aspirating and inserting the particles into PCR chamber, use a 1 mL pipette because it has large tip diameter enough to handle the particles. Tween-20 in rinsing buffer helps the particles not to adhere the surface of the pipette tip and

chamber wall. The particles can be located in the middle of the chamber by inserting rinsing buffer into the chamber mildly. After positioning, push the rinsing buffer with air using an empty pipette to remove the solution inside.

19. When inserting the PCR solution, be careful not to trap any air bubble in the chamber. Even small bubbles can be enlarged during PCR due to expansion of air in high temperature. Enlarged bubbles sometimes hinder the accurate fluorescence tracking because of fluorescence disturbance at the boundary of the bubbles.
20. We used two-step qPCR consisting of 8 s pre-denaturation, 3 s denaturation, and 11 s annealing steps. This protocol can be varied depending on the equipment and the premix.
21. Based on qPCR result without template, we assess whether the designed primers form primer-dimer in the particle or not. If any obvious signal is arisen, the primers should be redesigned to avoid it. By analyzing the graph of standard curve, the PCR efficiency in the particle can be calculated. The efficiency should be larger than 85%. If not, modulate the PCR conditions such as temperature and time in each step. In our experience, the primer-dimer formation is much more suppressed in particle-based qPCR than in solution-phase qPCR and PCR condition working well in solution-phase qPCR also does well in particle-based qPCR.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant No. HI13C2262). This work was also supported by KIST through the Institutional Program (Project No. 2E25590) and Open Research Program (Project No. 2E25722).

References

1. Meola N, Gennarino VA, Banfi S (2009) microRNAs and genetic diseases. PathoGenetics 2(1):7. doi:10.1186/1755-8417-2-7
2. Ono K, Kuwabara Y, Han JH (2011) MicroRNAs and cardiovascular diseases. FEBS J 278 (10):1619–1633. doi:10.1111/j.1742-4658.2011.08090.x
3. Roberts APE, Lewis AP, Jopling CL (2011) The role of MicroRNAs in viral infection. Prog Mol Biol Transl 102:101–139. doi:10.1016/B978-0-12-415795-8.00002-7
4. Nunez YO, Mayfield RD (2012) Understanding alcoholism through microRNA signatures in brains of human alcoholics. Front Genet 3:43. doi:10.3389/fgene.2012.00043
5. Deiuliis JA (2016) MicroRNAs as regulators of metabolic disease: pathophysiologic significance and emerging role as biomarkers and

therapeutics. Int J Obes 40(1):88–101. doi:10.1038/ijo.2015.170

6. Schmittgen TD, Lee EJ, Jiang J, Sarkar A, Yang L, Elton TS, Chen C (2008) Real-time PCR quantification of precursor and mature microRNA. Methods 44(1):31–38. doi:10.1016/j.ymeth.2007.09.006
7. Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung MW, Parsons JM, Steen MS, LaMadrid-Herrmannsfeldt MA, Williamson DW, Livingston RJ, Wu D, Wood BL, Rieder MJ, Robins H (2013) Using synthetic templates to design an unbiased multiplex PCR assay. Nat Commun 4:2680. doi:10.1038/ncomms3680
8. Jung S, Kim J, Lee DJ, Oh EH, Lim H, Kim KP, Choi N, Kim TS, Kim SK (2016) Extensible multiplex real-time PCR of MicroRNA using Microparticles. Sci Rep 6:22975. doi:10.1038/srep22975
9. Oh EH, Jung S, Kim WJ, Kim KP, Kim SK (2017) Microparticle-based RT-qPCR for highly selective rare mutation detection. Biosens Bioelectron 87(15):229–235. doi:10.1016/j.bios.2016.08.057
10. Choi NW, Kim J, Chapin SC, Duong T, Donohue E, Pandey P, Broom W, Hill WA, Doyle PS (2012) Multiplexed detection of mRNA using porosity-tuned hydrogel Microparticles. Anal Chem 84(21):9370–9378. doi:10.1021/ac302128u
11. Rutledge RG, Cote C (2003) Mathematics of quantitative kinetic PCR and the application of standard curves. Nucleic Acids Res 31(16):e93. doi:10.1093/nar/gng093
12. Durbeej B, Eriksson LA (2002) Reaction mechanism of thymine dimer formation in DNA induced by UV light. J Photoch Photobio A 152(1–3):95–101. doi:10.1016/S1010-6030(02)00180-6

Chapter 16

Optimized Whole Transcriptome Profiling of Motor Axons

Lena Saal-Bauernschubert, Michael Briese, and Michael Sendtner

Abstract

In highly polarized cells such as neurons, most RNA molecules are not randomly distributed but sorted into different compartments. So far, methods to analyze the transcriptome in distinct subcellular compartments are not well established. Here, we first describe the culturing of primary motoneurons in compartmentalized chambers to separate the axons from the somatodendritic compartment. Second, we introduce a method for whole transcriptome amplification followed by high-throughput sequencing to analyze the RNA composition of these two different compartments in neuronal cells.

Key words Compartmentalized cultures, Primary motoneurons, Axons, RNA, RNA-Seq, Transcriptome

1 Introduction

Within polarized cells the establishment of different subcellular compartments is thought to be guided, at least in part, by transport of specific transcripts into these subregions providing a basis for spatial and temporal control of protein synthesis [1]. Therefore, subcellular transcriptomic profiling has become increasingly important, especially in the field of neurobiology, as recent observations provide a link between axon guidance, regeneration, as well as presynaptic functions and local protein synthesis in the axon and axon terminal [2].

To analyze the axonal transcriptome neurons are typically grown in compartmentalized chambers to separately extract somatodendritic and axonal RNA for further processing and analysis. Since the amount of RNA which can be extracted from the axonal compartment is typically quite low, amplification steps need to be applied. Up to thousands of different axonal RNAs have already been detected by serial analysis of gene expression (SAGE) and microarray analysis [3–5]. However, current methods for transcriptome amplification are mostly based on oligo(dT)-based reverse transcription [6]. Therefore, to capture the whole subcellular

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_16,

transcriptome, including non-polyadenylated noncoding RNAs, we opted for a double-random priming strategy that can be utilized for the analysis of axonal transcriptomes. Hereby, we use an oligonucleotide which contains a random 3′ end for both reverse transcription and second strand synthesis followed by PCR amplification [7]. By applying this whole amplification method to RNA isolated from the axonal compartment of primary mouse motoneurons grown in compartmentalized chambers, we were able to quantitatively investigate very small amounts of RNA, including long noncoding RNAs and transcripts lacking polyadenylated 3′ ends in axons [8].

2 Materials

2.1 Compartmentalized Motoneuron Cultures

1. Motoneurons: Prepare according to the protocol by Wiese et al. [9].
2. Xona microfluidic chambers (150 μm).
3. Ultrasonic bath.
4. Washing solution for microfluidic chambers: Add 2 mL Micro-90® concentrated cleaning solution to 48 mL cell culture grade water. Always prepare fresh.
5. Nunclon 6 cm dishes.
6. Neurobasal medium supplemented with Glutamax: Mix 500 mL Neurobasal medium with 5 mL 100 × Glutamax. Store at 4 °C.
7. Borate solution: Prepare 150 mM boric acid in cell culture grade water and adjust pH to 8.35 (pH adjustment with NaOH). Autoclave buffer and store at room temperature.
8. Poly-DL-hydrobromide (PORN) solution: Dissolve 50 mg PORN in 1 mL 150 mM borate solution pH 8.35. Store 0.5 mL aliquots at −20 °C. Prepare the final solution of 0.5 mg/mL PORN in borate solution and filter through a 0.2 μm Corning filter. Store at −20 °C.
9. Laminin-111: Use a final concentration of 2.5 μg/mL in HBSS.
10. Full medium: For 50 mL mix 48 mL Neurobasal medium supplemented with Glutamax with 1 mL 50× B27 supplement and 1 mL heat inactivated horse serum and filter through a 0.2 μm Corning filter. Store at 4 °C.
11. Culture medium for somatodendritic compartment: Add recombinant rat CNTF (final concentration 5 ng/mL) to the appropriate volume of full medium needed. Always prepare medium fresh and make sure with adequate bioassays that this neurotrophic factor is fully biologically active.

12. Culture medium for axonal compartment: Add CNTF (final concentration 5 ng/mL) and BDNF (PeproTech) (final concentration 20 ng/mL) to the appropriate volume of full medium needed. Always prepare medium fresh.
13. PicoPure RNA Isolation Kit.

2.2 Whole Transcriptome Profiling

1. Water for molecular biology, certified RNase- and DNase-free: Use in all reactions and for primer dilution.
2. Primers: The original MALBAC primer sequence has been introduced by Zong et al. for the multiple annealing and looping-based amplification cycles (MALBAC) [10]. Order primers dissolved at 100 μM. Dilute MALBAC_primer to 50 μM with water. Prepare MALBAC_adapter mix containing MALBAC_adapter_1–4 at equimolar concentration and a total primer concentration of 50 μM (for example, 10 μL of each primer at a stock concentration of 100 μM and 40 μL water).

Primer name	Sequence (5′-3′)
MALBAC_primer	GTGAGTGATGGTTGAGGTAGTGTGGAGN NNNNNN
MALBAC_adapter_1	GTGAGTGATGGTTGAGGTAGTGTGGAG
MALBAC_adapter_2	GAGTGATGGTTGAGGTAGTGTGGAG
MALBAC_adapter_3	CTGTGAGTGATGGTTGAGGTAGTGTGGAG
MALBAC_adapter_4	TCTGTGAGTGATGGTTGAGGTAGTGTGGAG

3. QIAEX II Gel Extraction Kit.
4. DNA Marker for gel electrophoresis of PCR products: Dilute low molecular weight DNA ladder 1:30 with water. Take 5 μL, mix it with 1 μL 6 × DNA loading dye and load it onto the gel.
5. 10 × TBE buffer: Weigh 108 g Tris base, 55 g boric acid, and 9.3 g Na_2 EDTA. Add ultrapure water to 1 L. For final 1 × TBE gel running buffer dilute 50 mL of 10 × TBE with 450 mL ultrapure water.
6. 10% Ammonium persulfate (APS): Weigh 1 g and dissolve in ultrapure water to 10 mL.
7. Polyacrylamide gel: Mix 6.8 mL ultrapure water, 4 mL 30% acrylamide (29:1), 1.2 mL 10 × TBE buffer, 200 μL 10% APS, and 10 μL N,N,N′,N′-Tetramethylethylenediamine (TEMED). This gel solution is sufficient for two 10% Bio-Rad mini gels.
8. 0.1 × TE buffer: 1 mM Tris–HCl pH 7.4, 0.1 mM EDTA.
9. Primers for *Gapdh*: Forward 5′-GCAAATTCAACGGCACA-3′, Reverse 5′-CACCAGTAGACTCCACGAC-3′.
10. NEBNext Ultra DNA Library Prep Kit for Illumina.

3 Methods

3.1 Compartmentalized Motoneuron Cultures

1. Put your Xona microfluidic chamber into one 50 mL tube (*see* **Note 1**), add 50 mL washing solution and place the falcon into an ultrasonic bath for 30 min.
2. Exchange the washing solution with 50 mL cell culture grade water and place the falcon into an ultrasonic bath for another 30 min.
3. Discard the water and fill the falcon with 70% Ethanol. Incubate for at least 1 h.
4. Discard the Ethanol and gently place the chambers with the channels up into a dish under your cell culture hood (*see* **Note 2**). Let the chambers dry there overnight (*see* **Note 3**).
5. Put 2 mL PORN solution into a 6 cm dish and incubate the dish at 37 °C for at least 30 min.
6. Wash the dish three times with cell culture grade water and make sure to completely remove any water left (*see* **Note 4**).
7. Carefully take the microfluidic chamber with a sterile forceps and place it onto the PORN-coated dish with the channels down. Gently press with the back of the forceps onto the microfluidic chamber to make sure all air bubbles get removed, especially in the middle and at the edges (*see* **Note 5**).
8. Put 200 μL laminin into the upper left well. The solution will flow by capillary forces through the main channel and the microchannels as well (*see* **Note 6**).
9. Put 160 μL laminin into the lower left well and let the chamber stand for approximately 10 min (*see* **Note 7**).
10. Put 100 μL laminin solution into the upper right well of the microfluidic chamber and 60 μL laminin solution into the lower right well. Again try to avoid air bubbles.
11. Let the chambers with laminin solution stand at room temperature for the time of your preparation but at least for 1 h. Try to avoid any major temperature shifts during this time (*see* **Note 8**).
12. Take 1,000,000–1,300,000 motoneurons and centrifuge the cells at 200 × *g* for 5 min (*see* **Note 9**).
13. Carefully suck away the medium to the lowest volume possible (~10 μL) (*see* **Note 10**).
14. For virus transduction: Add the appropriate volume of virus at this point, bring the cell pellet into suspension, and let the cells incubate for 10 min at room temperature (*see* **Note 11**).
15. Take your prepared microfluidic chamber and wash all the channels containing the laminin as following. Carefully remove the laminin from all the wells and add Neurobasal medium in

the following order: upper left well 200 µL, lower left well 160 µL, upper right well 100 µL, lower right well 60 µL. Repeat this procedure once and again try to avoid air bubbles (*see* **Note 12**).

16. After the second wash remove all Neurobasal medium and add 200 µL full medium to all four wells.
17. For plating the cells suck away the full medium from both wells on the left side of the microfluidic chamber (*see* **Note 13**).
18. Bring your cell pellet in suspension with the remaining supernatant and plate ~3 µL of the cell suspension into the left main channel of the microfluidic chamber (*see* **Note 14**).
19. Repeat this procedure until no more cell suspension is left (*see* **Note 15**).
20. After all cells are plated into the microfluidic chamber let the chamber stand for approximately 5 min at room temperature to make sure all cells are settling down onto the laminin and are not washed out anymore.
21. Take the culture medium for the somatodendritic compartment and put 50 µL of it into the upper left well of the microfluidic chamber and directly another 50 µL into the lower left well.
22. Repeat this step once to come to a final volume of 100 µL in each well on the somatodendritic side (*see* **Note 16**).
23. Incubate the cells at 37 °C for 1 h.
24. Fill up each well on the somatodendritic side with 100 µL of the corresponding culture medium. On the axonal side of the chamber remove the medium in both wells completely and add 220–250 µL of the culture medium for the axonal side to each well (*see* **Note 17**).
25. Incubate the cells at 37 °C for the next 7 days. On day 1 and then every second day exchange half of the culture medium with fresh culture medium corresponding to the respective side.
26. Prior to RNA extraction wash the cells two times with prewarmed (37 °C) PBS. Remove the medium in the four wells and add PBS in the following order: upper left well 200 µL, lower left well 160 µL, upper right well 100 µL, lower right well 60 µL. Repeat this step once.
27. For RNA extraction remove the PBS from the two wells on the axonal side and put 100 µL extraction buffer from the PicoPure RNA Isolation Kit into the upper well. Wait approximately 10 s and collect all RNA extraction buffer from both wells. Repeat this procedure on the somatodendritic side (*see* **Note 18**).

28. Proceed with the RNA purification according to the manufacturer's instructions except that the isolated RNA is mixed with 100 µL 70% ethanol instead of 10 µL. Elute the purified RNA in 11 µL elution buffer and put on ice if used immediately or store at −80 °C.

3.2 Whole Transcriptome Profiling

1. For reverse transcription prepare the following reaction mix in a 0.2 mL PCR tube on ice:

Volume (µL)	Component
1	10 mM dNTPs
1	50 µM MALBAC_primer
1 or 10	RNA from the somatodendritic compartment or RNA from the axonal compartment
to 14.25	RNase- and DNase-free water
14.25	**Total**

2. Incubate the reaction mix at 65 °C for 5 min in a thermal cycler. Afterwards place immediately on ice and add the following components:

Volume (µL)	Component
4	5 × First-strand buffer
1	0.1 M DTT
0.25	RiboLock RNase Inhibitor (40 U/µL)
0.5	Superscript III Reverse Transcriptase (200 U/µL)
20	**Total**

3. Incubate the reaction mixture in a thermal cycler at 37 °C for 10 h, followed by an incubation at 70 °C for 15 min. Store the obtained cDNA at 4 °C (*see* **Note 19**).
4. For cDNA purification use the QIAEX II Gel Extraction Kit and proceed according to the manufacturer's instructions. Briefly, transfer the reverse transcription reaction mix to a 1.5 mL tube, add 60 µL buffer QX1, and mix. Add 10 µL QIAEX II suspension to each sample and proceed as described in the protocol. For elution of the cDNA resuspend the pellet in 20 µL water. After incubation and centrifugation transfer 19 µL of the supernatant into a new 0.2 mL tube. Take 1 µL of this supernatant, transfer it into a new 0.2 mL tube, and add 4 µL water. Store this 1:5 diluted sample at 4 °C as your "after RT" sample for quantitative PCR (*see* **Note 20**).

5. Take the leftover 18 μL supernatant containing the purified cDNA and add the following components for second strand synthesis:

Volume (μL)	Component
18	purified cDNA
1.725	50 μM MALBAC_primer
5	Accuprime buffer 2
1	Accuprime *Taq*
24.275	Water
50	**Total**

6. Incubate the reaction mixture in a thermal cycler at 98 °C for 5 min, followed by 37 °C for 2 min and 68 °C for 40 min.
7. For third strand purification use again the QIAEX II Gel Extraction Kit and transfer 19 μL of eluted third strand DNA into a new 0.2 mL tube.
8. For PCR amplification set up the following reaction mix:

Volume (μL)	Component
19	Purified third strand DNA
3.15	50 μM MALBAC_adapter mix
5	Accuprime buffer 2
1	Accuprime *Taq*
21.85	Water
50	**Total**

9. Incubate the reaction mix in a thermal cycler by using the following program:

Program step	Temperature (°C)	Time
1	92	2 min
x cycles of		
2	92	30 s
3	60	1 min
4	68	1 min
end		

10. Use 6 cycles for input RNA from the somatodendritic compartment and 18 cycles for input RNA from the axonal compartment.
11. **Optional:** After PCR amplification remove a 5 μL aliquot from each reaction and mix with 1 μL 6 × DNA loading dye. Load the DNA samples and marker on a 10% polyacrylamide gel and run them in 1 × TBE buffer for 25 min at 180 V. After the run is completed, disassemble the gel and incubate it in 50 mL 1 × TBE buffer containing 1 μL SYBR Green I for 5 min on a rocking platform. Wash the gel once with 1 × TBE buffer and view on a UV transilluminator (*see* **Note 21**).
12. After PCR amplification purify the PCR products with AMPure XP beads as following. Add 1.1× the volume of AMPure XP beads to the PCR products, mix gently by pipetting up and down, and incubate at room temperature for 5 min (*see* **Note 22**). Place the samples on a magnetic stand, wait for approximately 5 min and remove the supernatant whilst leaving the tubes on the stand (*see* **Note 23**). For washing the beads leave the samples on the magnetic stand and add 200 μL freshly prepared 80% ethanol. Wait for approximately 30 s, remove the ethanol, and repeat the procedure once for a total of two washes. After the second wash make sure to completely remove all the residual ethanol (*see* **Note 24**). Air-dry the beads for 10 min and add 50 μL 0.1 × TE buffer. Vortex briefly and incubate for 5 min. Put the samples back on the magnetic stand, wait for approximately 5 min, and collect 48 μL of the supernatant containing the purified PCR products (*see* **Note 25**).
13. After AMPure purification take 5 μL of each sample and subject these samples to gel electrophoresis as described under 11.
14. Furthermore, take 1 μL of each sample, dilute it 1:5 with water and subject it to quantitative PCR together with the "after RT" sample you stored at point 4. For quantitative PCR take primers recognizing *Gapdh* (*see* **Note 26**).
15. Additionally, you can measure approximate DNA concentration by using a Nanodrop (Peqlab).
16. For Illumina library generation take 50 ng of the AMPure purified PCR products from **step 12** and use the NEBNext Ultra DNA Library Kit for Illumina according to the manufacturer's instructions. Make sure to perform cleanup of adaptor-ligated DNA without size-selection and amplify the final library for eight cycles using the NEBNext High Fidelity 2 × PCR Master mix.
17. Again run 5 μL of each sample on a polyacrylamide gel as described for **step 11** (*see* **Note 27**).

18. Pool all your libraries by mixing together 10 μL of each library (*see* **Note 28**).
19. Purify the pooled libraries again using AMPure XP beads. Use the same volume of AMPure XP beads as the combined volume of your pooled libraries and perform purification as described for **step 12**.
20. Elute the pooled libraries in 50 μL 0.1 × TE buffer. They are now ready for sequencing.

4 Notes

1. You can put up to four microfluidic chambers into one 50 mL tube for washing but always make sure that they do not overlay. Otherwise a proper washing is not possible.
2. Don't take the chambers with your hand but always make sure to handle them with a sterile forceps.
3. The chambers need to be completely dry, otherwise they will be leaky later on.
4. The PORN plates also need to be completely dry, otherwise the chambers cannot adhere properly and will be leaky later on.
5. If air bubbles remain between the PORN dish and the microfluidic chamber, the chambers will be leaky later on. Also avoid removing the microfluidic chamber from the PORN plate after putting it there. This will destroy the PORN coating.
6. Try to avoid air bubbles in the main channel because they cannot be removed afterwards. Additionally, laminin coating will be incomplete at the position of air bubbles.
7. After 10 min you can check under the microscope if the laminin approaches the axonal side through the microchannels.
8. Once the microfluidic chambers are put onto the PORN dishes always store them at room temperature and never in the fridge. High temperature changes will make the chambers leaky.
9. Do not use more than 1,300,000 cells per chamber because the capacity of the main channel is limited. If you use less than 1,000,000 cells per chamber the cells will possibly not seed close enough to the microchannels and only few axons will approach the axonal compartment.
10. Not more than 10 μL of medium should be left, otherwise the cells will be too diluted and cell plating will be complicated.
11. Only virus preparations with high titer (~10^{10} PFU/mL) should be used to make sure that only a few μL of virus have to be added to the cells.

12. It is very important to wash the chambers with Neurobasal medium and to get rid of the laminin. Otherwise the laminin will clog the microchannels and the axons will not be able to grow through.
13. Try to put the tip of the suction pipette away from the entrance of the main channel, otherwise air bubbles can occur in the channel.
14. Place the tip containing the cell suspension directly in front of the entrance of the main channel and carefully empty the tip. The cell suspension will automatically flow into and through the main channel. Use 10 μL tips.
15. Do not plate the entire 10 μL of cell suspension at once because if the volume plated is too high the cells will float through the main channel and not stay there. Always plate the cells stepwise in small volumes.
16. Do not add 100 μL at once because this high volume of medium will wash out the cells from the main channel.
17. Always put a slightly higher volume of medium to the axonal side than to the somatodendritic side. This will ensure the establishment of the BDNF gradient.
18. The efficiency of RNA extraction can also be checked with a microscope. If the RNA extraction is successful, no intact cells or axons remain and the main channels appear completely empty.
19. Reverse transcription is done at 37 °C for 10 h to bring reactions to completion. Similar reaction conditions have previously also been used for cDNA synthesis from single cells (*see* ref. [11]).
20. It is important to remove an aliquot directly after reverse transcription to be able to check later on if amplification has worked.
21. Your PCR products should be visible on the gel as a smear sized 150–600 bp. Fragments <50 bp are second strand by-products.
22. For example, if your PCR reaction consists of a total volume of 50 μL, add 55 μL AMPure beads.
23. Wait until the solution is completely clear before removing the supernatant. The PCR products will remain on the beads.
24. For removing all residual ethanol, you can pulse spin your samples in a table-top microcentrifuge up to 1000 × *g* and remove trace ethanol with a 10 μL pipette.
25. Again make sure to remove the supernatant only if the solution is completely clear. Otherwise bead carry-over might occur and affect downstream applications.

26. As *Gapdh* is a house-keeping gene transcript levels are relatively abundant and therefore detection is quite robust. This control gives you a good estimation if the protocol was successful by a decrease in the crossing point of the amplified sample compared to the "after RT" sample.
27. The addition of the Illumina universal and index primer contributes an extra 122 bp to the library size. This becomes visible on your polyacrylamide gel as a respective shift of the size of the libraries compared to the PCR products from **step 11**.
28. If you pool all your libraries you only have one sequencing run. In total you can pool up to 12 different libraries in one sequencing sample.

Acknowledgments

This work was supported by Deutsche Forschungsgemeinschaft (BR4910/1-1 to M.B., SE697/4-1 to M.S.); Deutsche Forschungsgemeinschaft (BR4910/2-1 to M.B., SE697/5-1 to M.S.); Bayerischer Forschungsverbund Induzierte Pluripotente Stammzellen (D2-F2412.26 to M.S.); European Community's Health Seventh Framework Programme through the Euro-MOTOR Consortium (259867 to M.S.); and the Hermann und Lilly Schilling–Stiftung im Stifterverband der Deutschen Industrie.

References

1. Martin KC, Ephrussi A (2009) mRNA localization: gene expression in the spatial dimension. Cell 136(4):719–730
2. Lin AC, Holt CE (2008) Function and regulation of local axonal translation. Curr Opin Neurobiol 18(1):60–68
3. Andreassi C, Zimmermann C, Mitter R et al (2010) An NGF-responsive element targets myo-inositol monophosphatase-1 mRNA to sympathetic neuron axons. Nat Neurosci 13 (3):291–301
4. Gumy LF, Yeo GSH, Tung Y-CL et al (2011) Transcriptome analysis of embryonic and adult sensory axons reveals changes in mRNA repertoire localization. RNA 17(1):85–98
5. Saal L, Briese M, Kneitz S et al (2014) Subcellular transcriptome alterations in a cell culture model of spinal muscular atrophy point to widespread defects in axonal growth and presynaptic differentiation. RNA 20:1789–1802
6. Saliba AE, Westermann AJ, Gorski SA et al (2014) Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res 42 (14):8845–8860
7. Froussard P (1992) A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA. Nucleic Acids Res 20(11):2900
8. Briese M, Saal L, Appenzeller S et al (2016) Whole transcriptome profiling reveals the RNA content of motor axons. Nucleic Acids Res 44 (4):e33. doi:10.1093/nar/gkv1027
9. Wiese S, Herrmann T, Drepper C et al (2010) Isolation and enrichment of embryonic mouse motoneurons from the lumbar spinal cord of individual mouse embryos. Nat Protoc 5 (1):31–38
10. Zong C, Lu S, Chapman AR et al (2012) Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science 338(6114):1622–1626
11. Liss B (2002) Improved quantitative real-time RT-PCR for expression profiling of individual cells. Nucleic Acids Res 30(17):e89

Part IV

Protein Analysis

Chapter 17

2D-DIGE in Proteomics

Matias Pasquali, Tommaso Serchi, Sebastien Planchon, and Jenny Renaut

Abstract

The two-dimensional difference gel electrophoresis method is a valuable approach for proteomics. The method, using cyanine fluorescent dyes, allows the co-migration of multiple protein samples in the same gel and their simultaneous detection, thus reducing experimental and analytical time. 2D-DIGE, compared to traditional post-staining 2D-PAGE protocols (e.g., colloidal Coomassie or silver nitrate), provides faster and more reliable gel matching, limiting the impact of gel to gel variation, and allows also a good dynamic range for quantitative comparisons. By the use of internal standards, it is possible to normalize for experimental variations in spot intensities and gel patterns. Here we describe the experimental steps we follow in our routine 2D-DIGE procedure that we then apply to multiple biological questions.

Key words 2D-DIGE, Electrophoresis, SDS-PAGE, Proteomics, Isoelectrofocusing, Fungi, Plants, Animals

1 Introduction

Since 2D-DIGE technique was firstly presented [1], a technique based on the possibility to label lysine domains with different cyanine molecules, many discoveries and technical advances have brought to a standardized procedure that leads to proteomic advances in different biological domains. As all other 2D approaches, 2D-DIGE provides a map of proteins which reflects changes in both protein abundance levels and isoform variety. The main advantages of the technique are related to the fact that samples can be pooled; internal standards are present in every gel easing the normalization among runs and allowing automatic matching of spots among gels. Moreover, sensitivity is high (15% abundance change is robustly detectable) and the detection is linear above a 10,000-fold concentration range [2]. Sensitivity of new MS instruments is practically eliminating one of the drawbacks of the technique related to the identification process. The use of minimal labelling technique reduces the risk of multiple spots per protein

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_17,

due to multiple dye-molecule addition. By the technique described below we have generated maps and identified biological processes in many different domains confirming the usefulness of the technique in biology and medicine on human cells, plants, and fungi [3–5].

The aim of this protocol is to describe the procedure we follow for 2D-DIGE experiments focusing in details on the steps that are specific for 2D-DIGE (labelling and data acquisition). We suggest also valuable comprehensive reviews for 2D-PAGE in general [6–8].

2 Materials

All the procedures should be carried out using powder-free gloves (introducing fluorescent artifacts) and taking care to avoid any source of keratin contamination. Some of the procedures should be carried out under a chemical hood. All solutions should be prepared with ultrapure water and analytical grade level of reagents.

2.1 Labelling

1. Labelling buffer: 7 M Urea, 2 M Thiourea, 4% (w/v) CHAPS, Tris 30 mM. Weigh 42 g Urea, 15.2 g Thiourea, 4 g CHAPS, and 364 mg Tris. Dissolve in water for a final volume of 100 mL. Aliquot by 2 mL and store at −20 °C.
2. Labelling Stop solution: 10 mM Lysine. Dissolve 14.6 mg of lysine in a final volume of 10 mL with water. Aliquot by 50 μL and store at −20 °C.
3. Labelling CyDyes™: Add 25 μL dimethyl formamide (DMF, *see* **Note 1**) to 25 nmol of each dye (Cy2, Cy3, Cy5), store at −20 °C for a maximum of 3 months.
4. pH test paper for evaluating pH of solubilized proteins and 50 mM NaOH for pH adjustment.

2.2 Separation

1. IPG strips (GE Healthcare) with pH gradient that can vary depending on the specific research question.
2. Sample buffer (2×): 7 M urea, 2 M thiourea, 4% (w/v) CHAPS, 2% (v/v) ampholyte, 2% (w/v) DTT.
3. Rehydration tray, manifold, samples cups, paper wicks.
4. Precasted gels (e.g., Serva HPE™ horizontal gels depending on the separation system, *see* **Note 2**).
5. First and second dimension electrophoresis units. For second dimension we use HPE (high performance electrophoresis) device with a flat-bed 12.5% kit from Serva which include electrode buffers for cathode and anode, contact fluid, precast 2-D gel on a nonfluorescent plastic sheet. DL-DTT, iodoacetamide (IAA), urea, ice, microcentrifuge.

2.3 Data Acquisition and Analysis

1. Imaging device (laser scanner Typhoon FLA9500, GE Healthcare).
2. Software analysis (we use Decyder-GE Healthcare- and SameSpots-TotalLab).
3. Spot handling workstation.
4. Mass spectrometer.

3 Methods

The experimental design should be defined ahead considering that every gel can contain two samples as well as the internal standard. It is appropriate to apply the dye swap concept to avoid any dye effect, i.e., preferential labelling on the samples (Fig. 1). For statistical robustness of the results a minimum of four biological replicates is suggested. Internal standard is the result of a pooling process in equal amount of each sample of the analysis, usually labelled with Cy2 and used across the gels to allow a better matching as well as standardization.

Samples are prepared with the most suitable extraction method (*see* e.g., [9]) tested prior the final experimental setting leading to fine powder of the material under scrutiny. The quality of the extraction is the first factor affecting the overall results of the procedure; therefore it is suggested to carefully check the extracted material (*see* **Note 3**).

Dry protein pellet is then resuspended in the labelling buffer to a concentration of about 5 μg/μL. Resuspension occurs in a bath shaker at room temperature, or lightly heated (urea should not be heated over 37 °C) for 30–90 min depending on the characteristics of the pellet. Urea allows denaturation of all protein to a single conformation before running. After centrifugation (15,000 × *g*, 15 min), the supernatant is transferred in 1.5 mL tube. Verify that the sample pH lies in the range pH 8.0–9.0 (*see* **Note 4**).

Gel	Cy2	Cy3	Cy5
1	Internal std	Control	A
2	Internal std	Control	B
3	Internal std	A	Control
4	Internal std	B	Control
5	Internal std	B	A
6	Internal std	A	B

Fig. 1 Dye switch scheme. Each color represents an experimental condition [6], each letter a biological replicate. Note that each condition and each replicate are labelled two times with Cy3 and two times with Cy5

Prepare the internal standard by mixing the samples in equal amount (50 μg of total internal standard proteins will be loaded per gel).

3.1 Labelling of the Samples

1. Before labelling, IPG strips have to be hydrated with a solution containing ampholytes. Add 15 μL of ampholytes in 3 mL of Destreak rehydration solution, calculating 450 μL of this solution per 24 cm strip (350 μL for 18 cm; 200 μL for 11 cm).
2. Place the strips, with or without the plastic foil, in the Rehydration tray paying attention to avoid air bubbles). Then cover with paraffin oil in order to avoiding crystallization of urea and drying.
3. Let for rehydration at least 12 h, but not more than 18 h.
4. To start labelling add a volume of protein sample equivalent to 50 μg to a microcentrifuge tube. Add 0.4 μL of CyDye solution to the microfuge tube (*see* **Note 5**). The ratio of 400 pmol/μL of dye has to be respected to optimize labelling, mix dye and protein sample by vortexing. Then centrifuge briefly in a microcentrifuge to collect the solution at the bottom of the tube and incubate on ice for 30 min in the dark.
5. Add 1 μL of stop solution (*see* **Note 5**) to quench the reaction (for bulk labelling, add the drop on the side of the tube, like for the dye). Mix and spin briefly in a microcentrifuge and leave for at least 10 min on ice in the dark.
6. Combine the two or three differentially labelled samples into a single microfuge tube and mix. One of these samples should be the pooled internal standard (*see* **Note 6**).
7. Complete every tube containing Cy2, Cy3, Cy5 labelled samples to a final volume of 150 μL with sample buffer. The content of tube will then be loaded on a strip.

3.2 Separation

Cup loading of the IPG strips provides the best results in our lab as confirmed by comparative methods literature [10]. Although, it is recommended to make trials before starting the real experiment to choose for the best separation method.

3.2.1 Isoelectrofocusing

1. Place the manifold tray on the isoelectrofocusing (e.g., IPGphor, GE Healthcare) platform and verify the level.
2. Transfer the strips face up in the manifold, in the middle of the channel and put the (+) correctly oriented.
3. Place the Paper wicks, wetted with distilled water, such that they are overlapping the end of the strip. Then place the electrodes in contact with the wick. Then place the cups at ±1 cm of the end of the gel (cathodic side) and ensure that they are seated and not leaking (test with oil).

4. Centrifuge the sample and load it into the sample cups, then cover the sample with a few drops of paraffin oil and close the isoelectrofocusing platform (e.g., IPGphor) lid.
5. Select your program for the isoelectrofocusing and run accordingly (*see* **Note 7**).

3.2.2 Equilibration and Second Dimension

1. IPG strips are rinsed with distilled water and incubated in equilibration buffer containing 0.3% (w/v) urea and 0.8% DTT (w/v) for 15 min with gentle shaking and then a second step with the equilibration buffer complemented with 0.3% (w/v) urea and 2% (w/v) IAA for another 15 min (*see* **Note 8**).
2. To prepare HPE gels soak the electrode wicks from the kit (Serva kit of HPE) for 15 min with electrode buffer, one wick with anode buffer and one with cathode buffer per gel. Then 4 mL of cooling solution are added to the cooling plate. After setting the gel on the plate the cooling solution should form a homogeneous layer avoiding air bubbles. Then add the anode and cathode wicks. Then IPG strips are deposited, gel side down, in the strip-slot.
3. The run starts at 15 °C either for several hours or overnight. After 70 min of low-voltage steps (4 W for 30 min/12 W for 30 min/20 W for 10 min for four gels) the strip is removed from the gel (*see* **Note 9**). The run itself is set at 120 W for 4 h, then 160 W for 50 min for short run and 8 W overnight (10–15 h), and then 120 W for 3 h for overnight runs. The run is stopped when the dye front reaches the end of the gel (*see* **Note 10**).

3.3 Image Acquisition, Analysis and Spot Identification

3.3.1 Image Acquisition

1. Follow instructions of the adopted scanner and of the software used for image analysis.
2. Wipe carefully gel plastic backings and the scanner glass to avoid any dust (*see* **Note 11**).
3. Do prescan of the gel to assess the signal intensity using a low resolution image by scanning at three different lasers excitation wavelengths 488, 532, and 633 nm, acquiring light emitted at 520, 590, and 680 nm.
4. Adjust the PMT voltage setting for each channel to make sure that no saturation is present in the area of interest on the final image. Note: ideally the signal intensity should not vary more than 15% among the three images of the same gel (Fig. 2).
5. Do the final scanning increasing scanning resolution (to a pixel with a side of 100 μm) (*see* **Note 12**).
6. Perform a cropping procedure to eliminate unnecessary parts of the image.

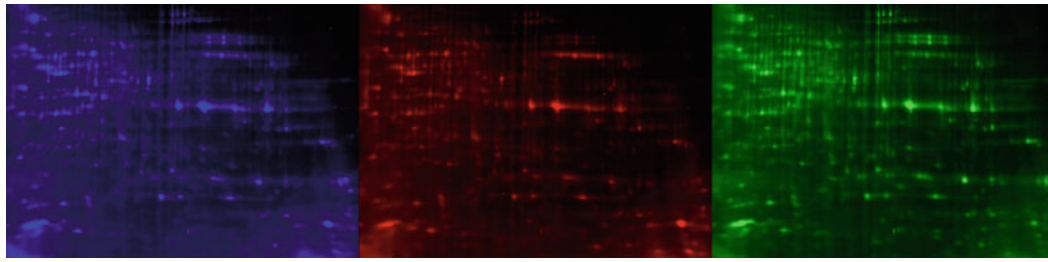


Fig. 2 Images of the same gel which include the three dyes and three different samples, in this example a whole fungal proteome from three strains [11]

3.3.2 Analysis and Identification

1. Image analysis can be carried out with different software (Image Master 2D platinum, SameSpots, Delta 2D, Decyder, etc.). In any case, follow the guidelines associated with the software. There are two basic processing types:

   A:

   (a) Load the images taking care of the orientation and cropping, i.e., the size of the gels should be similar across the experiment.

   (b) Detect spots and filter background with parameters such as spot volume, slope, area, and peak height.

   (c) Match the detected spots across all the gels based on the internal standards.

   (d) Select the spots of interest based on statistical tests (commonly *t*-test and/or ANOVA) realized on their normalized volumes.

   B:

   (a) Load the images taking care of the orientation and cropping.

   (b) Warp all images so that they are perfectly superimposed.

   (c) Detect spots on one image, transfer the spot boundaries to all other images, and filter background with parameters such as spot volume, slope, area, and peak height.

   (d) Select the spots of interest based on statistical tests (commonly *t*-test and/or ANOVA) realized on their normalized volumes.

2. Spots of interest can then be picked from one of the gels or from a preparative gel. Using a robot for the picking reduces risks of contamination and ensures a high accuracy of the picking (*see* **Note 13**).

3. Picked spots are then submitted to enzymatic digestion, usually with trypsin [11] although other enzymes can be used too. Resulting peptides are used for mass spectrometry analysis

using either MALDI TOF/TOF or ESI-MS/MS. Spectra are searched against protein databases. Generally, a protein is considered identified when at least two of its peptides have a significant score of identification. We do not rely anymore on peptide mass fingerprinting only.

4 Notes

1. After the reconstitution of the CyDye (25 μL DMF (less than 3 months old) to 25 nmol of dye), dyes can be kept at −20 °C for 3 months. DMF should be anhydrous (>99.8% purity to avoid degradation to amines of DMF caused by water). Put the dye in a light protecting Eppendorf. For the pipetting of the CyDyes we would recommend the use of low retention tips on a well calibrated P 2.5 pipette. The volume of the CyDyes is generally very small and the solution presents some viscosity. This will allow pipetting the entire volume in the vial without the any significant loss. If labelling 50 μg of proteins, it is advisable to dilute the dyes in DMF (from 1 ng/μL to 0.5 ng/μL); this would allow to pipet bigger volumes and have better accuracy. P 2.5 can only go down to 0.5 μL with an accuracy of 10%.
2. Gels can be manually casted but the use of pre-casted gels increases reproducibility of spot patterns between gels. A vertical apparatus can also be used.
3. We sometime verify on a small 2D gel the quality of the extract on a random sample. Protein quantification requires also care and precision as it influences the differential labelling procedure. Quantification is performed with kits based on precipitation of protein and colorimetric assay.
4. The pH of the extract can be adjusted to 8.5 by carefully adding diluted sodium hydroxide (50 mM). pH values outside this range will decrease the efficiency of the labelling reaction. The addition of 30 mM Tris to the labelling buffer normally ensures stability of the solution maintaining it in the correct pH range (if the sample is clean enough); however, this increases the amount of salts present in the strip and it may be required to extend the first step of the isoelectrofocusing at 30 V to ensure complete removal of all the salts prior to reaching high voltages. It is commonly admitted that IEF can stand a salt concentration of 50 mM.
5. For the labelling of a large number of samples, we recommend not to pipette the CyDyes and the stop solution directly into the labelling tubes. On the contrary, in our standard procedure, we deposit the drop of CyDye or of the Lysine (stop) solution on the side of the vial and we carefully position the vials in a

tabletop centrifuge. When all tubes are in the centrifuge, we spin and vortex the tubes and we start the incubation time. For stopping the labelling reaction, about 10 min before the end of the incubation we start depositing the drop of the Lysine solution and we spin and vortex when the timer for the incubation time goes off. This procedure allows uniform incubation times for all tubes.

6. The labelled samples can be processed immediately or stored for up to 3 months at −80 °C in the dark. For large experiments of labelling a small 2D gel can be run on a few samples for verifying the labelling efficiency.
7. During the run check the electrical current. In general, the system limits the energy supply when the resistance reaches 50 μA per strip (75 μA on recent systems) preventing overheating of the strip. This might be due to high salinity of the samples and might result in a poor resolution of the first dimension. To solve this issue, increase the time of the low voltage steps to improve desalting. The isoelectrofocusing can be performed using user-specific preferences. However, there are a number of basic rules to follow in order to get proper separation and avoid burning of the strips:

   Allow a sufficient time at low voltage. Two to three hours between 30 and 60 V should ensure removal of majority of dissolved salts, which are responsible for the increase in electrical current. Alternatively, desalting columns can also be used.

   Increase from low to high voltage in small steps, avoiding big jumps. We normally use 2–3 h gradients for each step.

   Allow the strip to rest at the new voltage for about 2 h before increasing further.

   Ensure that the strips are subjected to high voltage (10,000 V) for at least 6–8 h. This, together with total Vh parameter that should in any case be higher than 100,000 Vh, will ensure complete focusing of the high molecular weight proteins.

   A typical focusing program would look like this:

   (a) 3 h at 30 V.
   (b) 2 h gradient increase from 30 to 500 V.
   (c) 2 h at 500 V.
   (d) 2 h gradient increase from 500 to 1000 V.
   (e) 2 h at 1000 V.
   (f) 2 h gradient increase from 1000 to 3000 V.
   (g) 2 h at 3000 V.
   (h) 3 h gradient increase from 3000 to 5000 V.
   (i) 2 h at 5000 V.

(j) 2 h gradient increase from 5000 to 10,000 V.

(k) 8 h at 10,000 V.

8. Iodoacetamide alkylates the sulfhydryl groups of the proteins to prevent their potential reoxidation.
9. The immobilized pH gradient would interfere with the high voltage of the next steps resulting in a poor resolution.
10. Gels can be fixed prior scanning (>2 h in 15% Ethanol with 1% (w/v) citric acid). Gels between glass plates should be stored in air-tight box with wet paper at 4 °C, to prevent dehydration of the gel (note, the gels should not swim in water, this would trigger a diffusion of the protein spots). In such a case the gels can be kept 2–3 days.
11. Be careful not to use denatured ethanol, as they often contain fluorescent compounds that would increase the background of the images.
12. After acquisition, gels can be stored at 4 °C if they have been fixed. Gels can be stored for fairly a long time. We had experience of post-analysis performed 1 year after the separation. However, it is not possible to provide precise storage time. Long-term storage should be done in a solution containing a small amount of acid (1% citric acid) to avoid moulds growth.
13. A preparative gel, loaded with a larger amount of protein coming from the different samples, can be prepared. This increases the chances of identification. The preparative gel has to be matched with the analytical gels or with the reference gel. An increased amount of protein often impacts the 2D pattern and can complicate the matching. Be careful not to overload the gel capacity. Usually, 500 μg of proteins are sufficient.

References

1. Unlü M, Morgan ME, Minden JS (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. Electrophoresis 18:2071–2077. doi:10.1002/elps.1150181133
2. Viswanathan S, Ünlü M, Minden JS (2006) Two-dimensional difference gel electrophoresis. Nat Protoc 1:1351–1358. doi:10.1038/nprot.2006.234
3. Bertrand A, Bipfubusa M, Castonguay Y et al (2016) A proteome analysis of freezing tolerance in red clover (Trifolium pratense L.) BMC Plant Biol 16:65. doi:10.1186/s12870-016-0751-2
4. Miller I, Diepenbroek C, Rijntjes E et al (2016) Gender specific differences in the liver proteome of rats exposed to short term and low-concentration hexabromocyclododecane (HBCD). Toxicol Res 5:1273–1283. doi:10.1039/c6tx00166a
5. Pasquali M, Serchi T, Cocco E et al (2016) A Fusarium graminearum strain-comparative proteomic approach identifies regulatory changes triggered by agmatine. J Proteome 137:107–116. doi:10.1016/j.jprot.2015.11.010
6. Görg A, Weiss W, Dunn MJ (2004) Current two-dimensional electrophoresis technology for proteomics. Proteomics 4:3665–3685. doi:10.1002/pmic.200401031
7. López JL (2007) Two-dimensional electrophoresis in proteome expression analysis. J Chromatogr B 849:190–202. doi:10.1016/j.jchromb.2006.11.049

8. Rabilloud T, Lelong C (2011) Two-dimensional gel electrophoresis in proteomics: a tutorial. J Proteome 74:1829–1841. doi:10.1016/j.jprot.2011.05.040
9. Pasquali M, Giraud F, Lasserre JP, Planchon S, Hoffmann L, Bohn T et al (2010) Toxin induction and protein extraction from *Fusarium* spp. cultures for proteomic studies. J Vis Exp 36:1690. doi:10.3791/1690
10. Dépagne J, Chevalier F (2012) Technical updates to basic proteins focalization using IPG strips. Proteome Sci 10:54. doi:10.1186/1477-5956-10-54
11. Pasquali M, Serchi T, Renaut J et al (2013) 2D difference gel electrophoresis reference map of a Fusarium graminearum nivalenol producing strain. Electrophoresis 34:505–509. doi:10.1002/elps.201200256

Chapter 18

STAGE-Diging in Proteomics

Paolo Soffientini and Angela Bachi

Abstract

Proteomics is nowadays a standard tool in life sciences for the analysis of protein abundance, modifications and interactions but has so far failed to enter the clinic for routine applications. New generation mass spectrometers and chromatographic systems are able to cover approximately an entire cell proteome in one run but sample preparation, in terms of time and sample recovery, is still a critical step. Here we describe a modification of the in-gel digestion method, called STAGE-diging, that reduces sample handling, decreases the analysis time and improves protein identification and quantification. This method is particularly useful for those research labs that manage different biological samples and have a limited access to MS instrumentation or are required to perform high-throughput analysis in short time like a clinical laboratory.

Key words Proteomics, In-gel digestion, Mass spectrometry, High-throughput, STAGE-diging, Sample processing

1 Introduction

In the last 20 years, proteomics evolved enormously in the fields of analysis of protein abundance, detection of post-translational modifications, and protein–protein interactions. Technological advancements of high-resolution, new generation mass spectrometers combined with Ultra Performance Liquid Chromatography [1, 2] made high-throughput proteomics available in many laboratories. Sample preparation instead, and in particular, digestion of proteins into peptides, that is the prerequisite of MS-based proteomics bottom-up approach, still relies on very established protocols. While for in-solution digestion of protein samples and cell lysates new protocols have been recently proposed [3–6], in-gel digestion procedures are still based on the protocol proposed more than two decades ago, with the exception of slight modifications [7–10].

Advantages of in-gel digestion that makes it preferable to other sample preparations are that it is a simple and cost-effective procedure for sample pre-fractionation and it has the ability to remove contaminants and detergents that can interfere with digestion and

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_18,

MS analysis. Moreover, in-gel digestion provides visual quality control of the samples as it can assess proteins mixture complexity and abundance and it is a highly efficient denaturation method that can be applied to a large variety of sample types [11]. Disadvantages that make the technique relatively low throughput are the requirement of a rather laborious process with numerous steps of washing and incubation that are still operator dependent, and the lower enzymatic efficiency (~20%) relative to that in solution that produce a lower efficiency in peptides recovery and predisposes samples to stochastic mistakes and contaminations.

Here, we present a faster and highly reproducible adaptation of the in-gel digestion method called STAGE-diging [12] where an entire gel lane is processed in a single, enclosed stop-and-go extraction tips (StageTip) [13]. This procedure can be applied both on high and low complexity samples and in proteomics qualitative and quantitative studies, with a consistent saving of time both in sample processing and in MS analysis time because an entire lane can be analyzed in a single run.

2 Materials

2.1 Stage Tip Assembly

1. Empore reversed-phase extraction disks from 3 M (C18 (ODS or Octadecyl) reversed-phase material, 3 M product number 2215).
2. 18 gauge blunt ended syringe needle.
3. p1000 pipette tips (Gilson or similar).
4. 0.3 or 0.5 μm ID (PEEK or fused silica) tubing.
5. Activation solution: 100% Methanol.
6. Conditioning solution: 0.1% Formic Acid (FA).

2.2 In-Gel Digestion

Prepare all solutions using UHQ (Ultra High Quality) water obtained from a Milli-Q system and analytical grade reagents.

1. Digestion buffer: 100 mM ammonium bicarbonate pH 7.8 in UHQ water.
2. Reduction buffer: 10 mM dithiothreitol in the digestion buffer.
3. Alkylation buffer: 55 mM iodoacetamide in the digestion buffer.
4. Trypsin stock solution: Trypsin 0.1 μg/μL in 1 mM HCl. Working solution: 12.5 ng/μL in digestion buffer.
5. Acetonitrile HPLC-MS grade 100%.
6. Elution solution: 80% ACN, 0.1% FA.
7. Sample buffer: 2% ACN, 0.1% FA (generally buffer A in reverse phase chromatography).

3 Methods

3.1 Double Plug p1000 STAGE-Diging Assembly

The Stage tip workflow is adapted from the method described by Rappsilber et al. [13].

1. Place an Empore disk on a clean hard surface, for instance a glass microscope slide or a Petri dish.
2. Press the 18 gauge blunt ended syringe needle into the Empore disk to core out a piece of C18 filter material.
3. Place the needle into a p1000 pipette tip and push the cored disk pieces into the pipette tip with the help of a PEEK or fused silica tubing. Gently pack the material into the end of the pipette tip (*see* **Note 1**).
4. Press a second C18 plug into the syringe needle for extra loading capacity and repeat **step 3** (*see* **Note 2**).

3.2 STAGE-Diging Protocol

The digestion workflow was adapted from the method described by Shevchenko et al. [8] and was tested both on Coomassie blue or Silver [9] stained gels. A brief summary of the workflow is shown in Fig. 1.

1. Cut an entire SDS-page lane or portion of interest of the gel into ~1 mm^3 cubes and transfer it into the STAGE-diging tip.
2. Dehydrate with 200 µL of 100% acetonitrile (ACN) for 3 min and then remove the solution by centrifugation using the commercial tip box as holder (*see* **Notes 3**–**5**).
3. Rehydrate the gel cubes with 200 µL digestion buffer for 3 min and remove the solution by centrifugation.
4. Repeat **steps 2** and **3** twice before dehydrating the sample by the addition of 200 µL of ACN for 3 min and subsequent centrifugation.
5. Reduction of protein disulfide bonds is carried out with 200 µL of reduction buffer, 30 min at room temperature and then the reduction solution is removed by centrifugation. Gel cubes are subsequently dehydrated with 200 µL of 100% ACN for 3 min and then submitted to centrifugation.
6. Alkylation of reduced cysteines is performed with 200 µL of alkylation buffer at room temperature for 30 min in complete darkness, then the solution is removed by centrifugation.
7. Rehydrate gel pieces for 3 min in 200 µL of digestion buffer, centrifuge and dehydrate with 200 µL of ACN for 3 min and then resubmit to centrifugation.
8. Rehydrate gel cubes with 40 µL of Trypsin (working solution) and after few minutes add 160 µL of digestion buffer and incubate at 37 °C overnight (or 6 h) in a commercial tip box

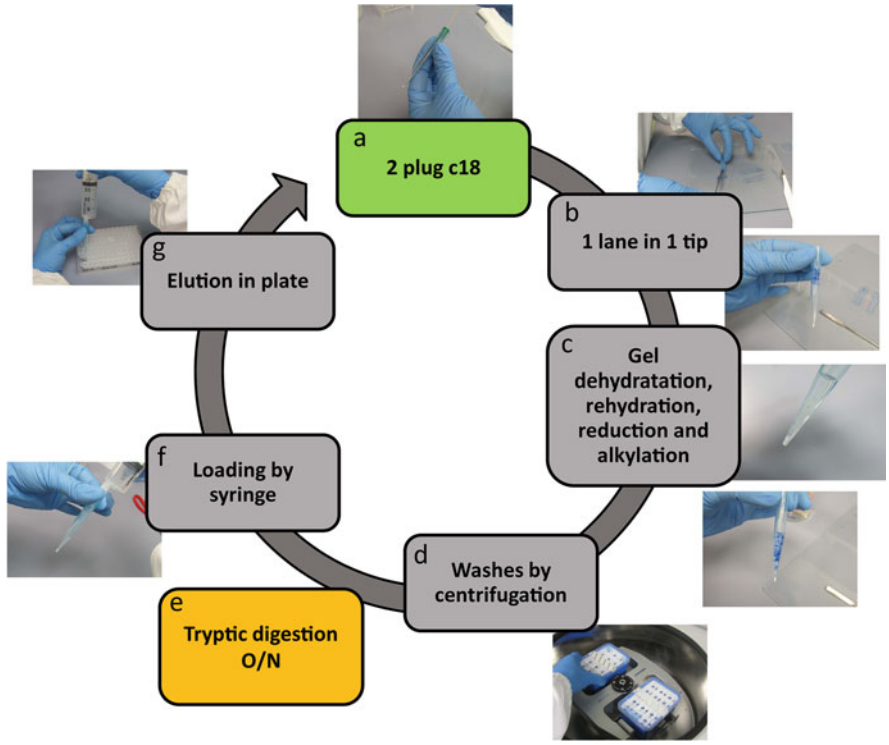


Fig. 1 STAGE-diging workflow. To enable the loading of 8–10 μg of total proteins a double C18 plug (**a**) is used. An entire gel lane, cut in small gel cubes, is transferred in the device (**b**). All the standard processes of dehydration of gel pieces, rehydration, reduction, and alkylation are performed sequentially (**c**) and removal of solutions is accomplished by centrifugation using the commercial tip box as holder (**d**). Alternatively, the solutions can be forced through the double C18 plug by pushing with a syringe. Enzymatic digestion is performed in the same commercial tip box, filled with water on the bottom to avoid buffer evaporation (**e**). The digestion solution is then forced through the device with a syringe to allow the loading of peptides onto the C18 plugs (**f**). The peptides are desalted with a washing step and then eluted two times directly in a 96 wells plate (**g**)

filled by water on the bottom to ensure that the buffer will not evaporate.

9. After protein digestion, force the solution through the STAGE-diging with a syringe.
10. Acidify the sample with 200 μL of conditioning solution for 3 min and then force this solution out with the syringe. In this way desalting of peptides can occur.
11. Elute peptides twice by adding 100 μL of elution solution. Push this solution through the double plug with the syringe (*see* **Note 6**).
12. Dry in a Speed-Vac the eluate and suspend it in 20 μL of sample buffer. The sample is now ready to be analyzed by LC-MS (*see* **Note 7**).

4 Notes

1. Do not over pack or under pack the C18 plugs in the p1000 tip, just push them gently.
2. Estimation of binding capacity per double plug of C18 core is 8–10 μg. Take this into account when you load gels and you choose to process them by STAGE-diging protocol.
3. To ensure that the gel pieces do not create a sticky surface on the C18, all the solutions can be added with a gel-loader tip.
4. All centrifugation are performed in Eppendorf 5810 equipped with A-2-DWP rotor, at 1800 rpm (532 rcf).
5. If removal of solutions cannot be accomplished by centrifugation, solutions can be forced through the double C18 plug by pushing with a syringe.
6. After the first elution step of peptides gel cubes will be shrank and they should appear of white color.
7. Both LC gradient slope and length and MS acquisition method must be adjusted according to the complexity of the sample. For few nanograms of samples a shorter gradient/acquisition method is preferred while for complex samples (micrograms scale) a longer gradient are preferred. The time saved by the possibility of injecting an entire lane in a single run allows performing technical replicates to increase the amount of peptides and protein identified and quantified.

References

1. Abersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422 (6928):198–207
2. Hebert AS, Richards AL, Bailey DJ, Ulbrich A, Coughlin EE, Westphall MS, Coon JJ (2014) The one hour yeast proteome. Mol Cell Proteomics 13(1):339–347
3. Wiśniewski JR, Zougman A, Nagaraj N, Mann M (2009) Universal sample preparation method for proteome analysis. Nat Methods 6 (5):359–362
4. Manza LL, Stamer SL, Ham AJ, Codreanu SG, Liebler DC (2005) Sample preparation and digestion for proteomic analyses using spin filters. Proteomics 5(7):1742–1745
5. Gobom J, Nordhoff E, Ekman R, Roepstorff P (1997) Rapid micro-scale proteolysis of proteins for MALDI-MS peptide mapping using immobilized trypsin. Int J Mass Spectrom Ion Process 199(169–170):153–163
6. Kulak NA, Pichler G, Paron I, Nagaraj N, Mann M (2014) Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. Nat Methods 11(3):319–324
7. Rosenfeld J, Capdevielle J, Guillemot JC, Ferrara P (1992) In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. Anal Biochem 203(1):173–179
8. Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat Protoc 1(6):2856–2860
9. Shevchenko A, Wilm M, Vorm O, Mann M (1996) Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. Anal Chem 68(5):850–858
10. Anjo SI, Santa C, Manadas B (2015) Short GeLC-SWATH: a fast and reliable quantitative approach for proteomic screenings. Proteomics 15(4):757–762
11. Switzar L, Giera M, Niessen WM (2013) Protein digestion: an overview of the available

techniques and recent developments. J Proteome Res 12(3):1067–1077

12. Soffientini P, Bachi A (2016) STAGE-diging: a novel in-gel digestion processing for proteomics samples. J Proteome 17(140):48–54
13. Rappsilber J, Ishihama Y, Mann M (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal Chem 75(3):663–670

Chapter 19

Protein Arrays I: Antibody Arrays

Yulin Yuan, Zuan-Tao Lin, Hongting Wang, Xia Hong, Mikala Heon, and Tianfu Wu

Abstract

Antibody arrays represent one of the very early protein array systems where antibodies are used to capture and detect target proteins in a high-throughput platform. The development of high-quality antibodies, nanomaterial-based novel detection probes, as well as innovative imaging technologies and computational tools has tremendously improved the sensitivity, specificity, and robustness of antibody arrays during the past decade. In this protocol we will incorporate the most updated innovations and developments of antibody arrays into the step-by-step experimental procedures. This includes antibody printing, sample preparation, array detection, as well as imaging and data analysis. Antibody array could be used for cytokine profiling or mapping of phosphorylation, glycosylation, or other post-translational modifications of target proteins.

Key words Antibody array, Cytokine profiling, Kinome mapping, Glycome mapping, Biomarkers

1 Introduction

Antibody array is an antibody-based high-throughput platform for protein profiling, screening, and comparison between an experimental group and a control group [1]. An attractive feature of antibody array technology lies in its capability of profiling proteins in non-fractionated biological samples, detecting a wide concentration range of analytes in a high-throughput and multiplex fashion [2]. As illustrated in Fig. 1, there are two strategies for protein detection using antibody arrays: (1) single-antibody method, where biotinylated protein samples are captured, followed by the detection using streptavidin-conjugated fluorescent dye; (2) sandwich-complex method, where a pair of antibodies (capture antibody and a biotinylated detection antibody) against the target protein are used, followed by the detection using streptavidin-conjugated fluorescent dye. The single-antibody method is simple and straightforward requiring only the capture antibody for the array. However, the obvious disadvantages of single-antibody method include the

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_19,

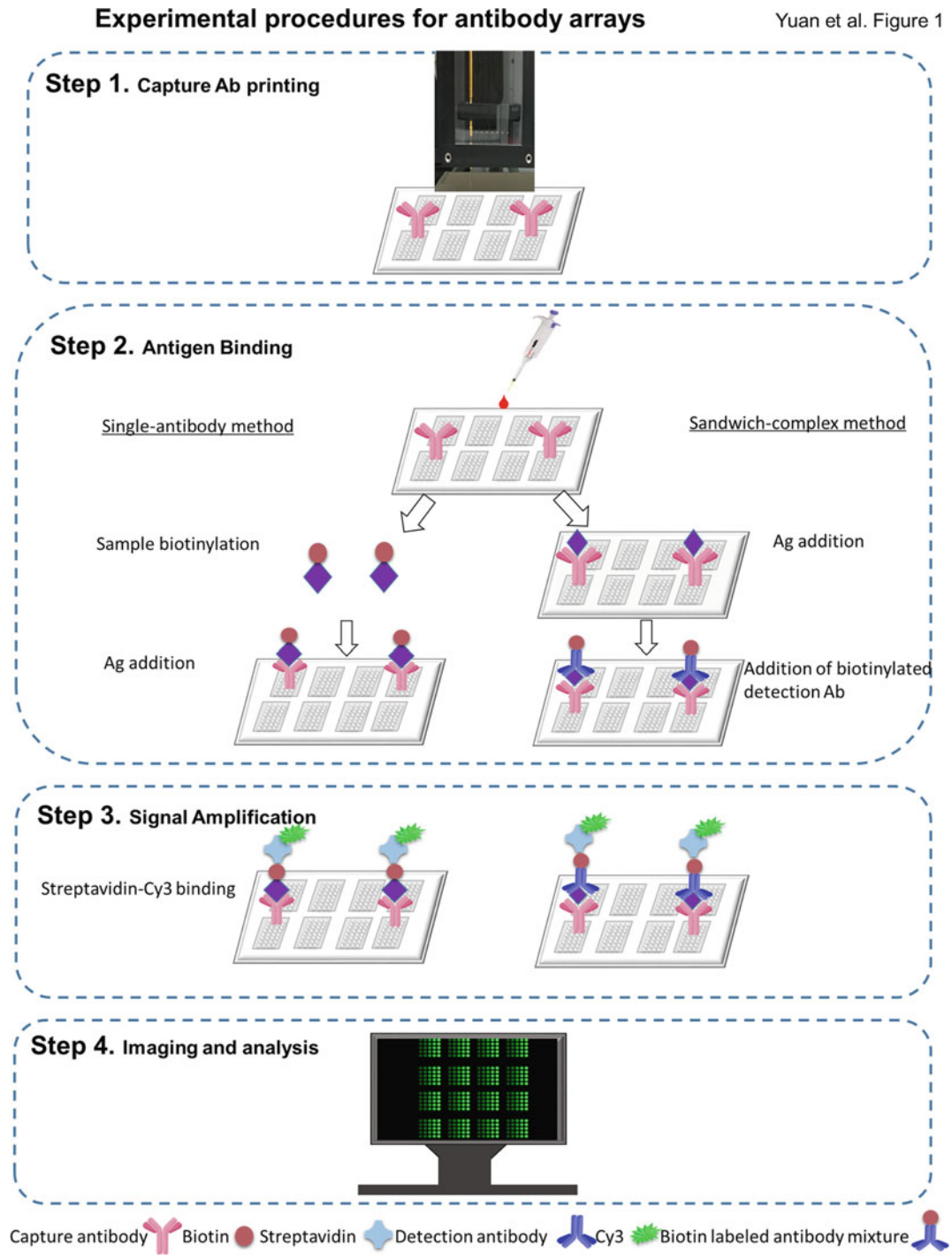


Fig. 1 Workflow of antibody arrays. Upon the printing of a panel of capture antibodies on glass slides, there are two strategies for the detection of the array: the single-antibody method and the sandwich-complex method. Shared procedures of the two methods are in *bold*. *Ab* antibody, *Ag* antigen

compromised specificity, increased background signals, and decreased sensitivity as compared to sandwich-complex method. The sandwich-complex method requires a pair of antibodies per antigen, each recognizing a different epitope within the protein. However, this limits the number of proteins to be analyzed due to the commercial availability of paired antibodies.

2 Materials

Milli-Q water was used throughout.

PBS: 137 mM NaCl, 2.7 mM KCl, 4.3 mM Na_2HPO_4, 1.47 mM KH_2PO_4.

PBST: 137 mM NaCl, 2.7 mM KCl, 4.3 mM Na_2HPO_4, 1.47 mM KH_2PO_4, 0.1% Tween-20, pH 7.4.

2.1 Preparation of Capture Antibodies

1. Capture antibodies: From commercial sources. Store the antibody stock solutions at 4 °C or −20 °C according to the manufacturer's instructions (*see* **Note 1**).
2. Antibody dilution buffer: PBST.

2.2 Antibody Printing

1. Printing buffer: PBS.
2. Slide: Sony DADC Epoxy MS slide (STRATEC Consumables GmbH, Austria) (*see* **Note 2**).
3. SciFlexarrayer S3 (Scienion, Berlin, Germany) (*see* **Note 3**).
4. Positive control: Pierce™ Bovine Serum Albumin, Biotinylated (Thermo Fisher Scientific Inc., USA) (*see* **Note 4**).
5. Negative control: PBS solution (*see* **Note 5**).
6. 384-Well microarray microplates (Arrayit Corporation, USA).

2.3 Sample Preparation

1. Serum from donors (*see* **Note 6**).
2. Pierce™ BCA Protein Assay Kit (Thermo Scientific™, USA).
3. Sample dilution buffer: 1% BSA in PBST or 3% nonfat milk in PBST (PBS-MT), or Super G blocking buffer (Grace bio-labs, Inc., USA).

2.4 Labeling of Samples (For Single-Antibody Method)

Sample labeling reagent: EZ-Link™ Sulfo-NHS-LC-Biotin (Thermo Fisher Scientific Inc., USA).

2.5 Array Detection

1. Washing buffer: PBST.
2. Blocking buffer: The same as sample dilution buffer.
3. Detection antibodies: Biotinylated antibodies.
4. Cy3-conjugated streptavidin (Thermo Fisher Scientific Inc., USA).

2.6 Data Analysis

1. Scanner: GenePix Microarray Scanner 4400A (Molecular Devices LLC, USA) (*see* **Note 7**).
2. Analysis software: GenePix Pro 7 software (Molecular Devices LLC, USA) (*see* **Note 8**).

3 Methods

To avoid contamination, wear gloves while performing the procedures.

3.1 Preparation of Antibodies

1. Antibodies are retrieved from freezer or refrigerator (*see* **Note 9**).
2. Serially dilute capture antibodies and detection antibodies with printing buffer to acquire working concentrations (*see* **Note 10**).

3.2 Printing of Antibody Microarrays

1. Spin the tubes containing diluted antibodies and controls.
2. Transfer the antibodies into a 384-well microplate with 40 μl per well.
3. Cover the slide with Microseal® 'B' Adhesive Seals and centrifuge the microplate 2000 × *g* for 5 min at 4 °C to remove any bubbles.
4. Fill the wash bottle of SciFlexarrayer S3 with millipore water and empty the waste bottle.
5. Set the relative humidity to 50% and ambient temperature to 4 °C in the printing chamber.
6. Turn on the SciFlexarrayer S3 and open the software.
7. Choose the "prime" option to initiate a standard starting routine.
8. Click "Do Task," then "sciCLEANWash Tray" to wash the Piezo Dispense Capillary (PDC).
9. Mount Epoxy slides onto the SciFlexarrayer S3.
10. Target setup
    (a) Target: Select the fields where samples will be printed.
    (b) Field setup: According to the number of antibodies that need to be printed to set up the work table; A twofold dilution gradient will be set up for each antibody to find an optimal concentration (*see* **Note 11**).
11. Click "Run" to start printing.
12. When printing is finished, wash the PDC three times with 1% sciCLEAN 8.
13. Use the slides after air drying for 20 min, or store them at 4 °C (*see* **Notes 12** and **13**).
14. Cover the 384-well microarray microplates with Microseal® 'B' Adhesive Seals and store the plates in −80 °C (*see* **Note 14**).

3.3 Blocking

1. Place the 16-well Slide Modules on the bench and cover the silicone layer.
2. Carefully place the glass slide face down to the silicone layer gasket.
3. Insert the metal clip into the groove in the gasket and rotate the clip into the locked position.
4. Slide the clip into place. Do the same for the other side.
5. Add 100 μl blocking buffer to each sub-array.
6. Cover the incubation chamber with ProPlate® Slide Module Seal Strips and incubate the slides at room temperature for 60 min on an orbital shaker rotating at 55 rpm.
7. Aspirate blocking buffer from each well.
8. Add 100 μl/well washing buffer, then incubate on the orbital shaker rotating at room temperature for 5 min (55 rpm). Repeat this step twice, aspirating between washes (*see* **Note 15**).

3.4 Sample Preparation

1. Collect whole blood in a red-topped Vacutainer.
2. Leave the blood undisturbed at room temperature for 30 min to clot.
3. Centrifuge the tube at 2000 × *g* for 10 min at 4 °C.
4. Determine the total protein concentration of the analytes using Pierce™ BCA Protein Assay Kit.
5. According to the protein concentration, dilute the sample to the appropriate concentration.

3.5 Protein Labeling (This Step Is for Single-Antibody Method Only!)

1. Biotin preparation: According to the manufacturer's instructions.
2. Aliquot samples to 20 μl/tube at the concentration of 2.5 μg/μl.
3. Add labeling buffer to bring the volume to 100 μl.
4. Add 4 μl of biotin/DMSO solution. Incubate the mixture at room temperature for 1 h, shaking every 10 min.
5. Add 50 μl of Stop reagent. Mix the reaction reagents by turning upside down. Incubate at room temperature for 30 min, shaking every 10 min.
6. Use the biotinylated sample immediately or store it at −80 °C (*see* **Note 16**).

3.6 Array Detection

3.6.1 Sandwich-Complex Method

1. Dilute samples to 50- to 200-fold using sample dilution buffer (*see* **Note 17**).
2. Add 75 μl of diluted sample to each sub-array.
3. Cover the ProPlate® 16 Well Slide Module with ProPlate® Slide Module Seal Strips.

4. Place the arrays into a humidified chamber with 100% humidity and incubate at room temperature with shaking (55 rpm) for 60 min.
5. Remove ProPlate® Slide Module Seal Strips and carefully aspirate samples from the sub-arrays, touching only the corners with the pipette tips.
6. Add 100 μl/well washing buffer and incubate on an orbital shaker rotating at 55 rpm for 5 min at room temperature. Repeat this step four times.
7. Incubate with 100 μl/well biotin-labeled antibody cocktail with shaking (55 rpm) for 60 min at room temperature.
8. Wash the arrays using 100 μl/well washing buffer (5 × 5 min) with shaking (55 rpm).
9. Incubate with 100 μl/well Cy3 coated streptavidin in blocking buffer with shaking (55 rpm) for 60 min at room temperature (*see* **Note 18**).
10. Wash the arrays using 100 μl/well wash buffer (5 × 5 min) with shaking (55 rpm).
11. Wash the arrays using 100 μl/well PBS (3 × 5 min) with shaking (55 rpm).
12. Wash the arrays using 100 μl/well H_2O for 5 min with shaking (55 rpm).
13. Remove slides from ProPlate® 16 Well Slide Module and immerse the entire slides into washing buffer in a 50 ml conical tube, shake at room temperature for 5 min (55 rpm).
14. Discard the washing buffer and add 45 ml H_2O, shake at room temperature for 5 min (55 rpm).
15. Discard the liquid and centrifuge the conical tube at 2000 × *g* for 2 min.
16. Place the slides into a new conical tube to air dry or under a stream of nitrogen gas.
17. Scan or store the slide at room temperature in a dark chamber.

3.6.2 Single-Antibody Method

1. Add 75 μl/well biotin-labeled samples into sub-array and incubate for 1 h.
2. Cover the ProPlate® 16 Well Slide Module with ProPlate® Slide Module Seal Strips.
3. Place the arrays into a humidified chamber with 100% humidity and incubate at room temperature with shaking (55 rpm) for 60 min.
4. Remove ProPlate® Slide Module Seal Strips and carefully aspirate samples from the sub-arrays, touching only the corners with the pipette tips.
5. Repeat the **steps 8–17** of Subheading 3.6.1.

3.7 Scan and Data Analysis

1. Turn on the GenePix 4400A scanner and warm up for 20 min.
2. Insert dry slides upside down into the holder of the scanner.
3. Perform a preview scan of the entire slide.
4. According to the preview result, set a suitable laser power, PMT gain, and scan area (*see* **Note 19**).
5. Scan the slide with the selected settings and save the image.
6. Open the images of the microarrays using GenePix Pro 7 software and load the array list (GAL file).
7. Adjust the brightness and contrast (*see* **Note 20**).
8. Normalize chips based on the positive control, adjusting the signal intensities of the control to the same value in all the datasets to be compared (*see* **Note 21**).
9. Assess the sensitivity of the antibody and the linearity of the detected signal by a dilution series of each sample to build a dilution curve (*see* **Note 22**).
10. Calculate the average of the duplicate spots to obtain a value for the well (array). Subtract the local background and calculate the signal-to-noise ratio (SNR) (*see* **Note 23**).
11. The signal intensity values represent the protein expression level.

4 Notes

1. Several types of antibodies including monoclonal, polyclonal, recombinant antibody fragments or single-chain Fv (scFv) antibodies can be used as capture antibodies on the arrays.
2. A wide variety of surfaces can be used as solid supports for antibody microarrays, most commercial supports have been successful in our hands, such as Arrayit® Corp (SuperNitro coating), GE Healthcare (FAST® slide), Grace Bio-Labs (ONCYTE® Avid™ or Nova™ film slides), or Schott AG (Nexterion® C or NC slides).
3. The selection of a proper printer is based on sample type and printing speed [3]. We use a non-contact printer SciFlexarrayer S3 equipped with specifically coated nozzles according to sample type.
4. Different reagents (proteins or dyes) can be used as positive controls. We have been using biotinylated BSA as a positive control.
5. Unrelated antibodies, isotype controls, or simply spotting buffer could serve as negative controls. In our studies, we often use spotting buffer as a negative control.

6. The samples used in an antibody array could include cell lysates, fresh or frozen tissue lysates, body fluids such as serum, urine, tear, sweat, or synovial fluids, etc. [4, 5]. The lysates or serum should be aliquoted and rapidly frozen at −20 or −80 °C. Avoid freeze-thaw cycles.
7. Different types of scanners can be used for the imaging according to staining method. We use the GenePix Microarray Scanner 4400A for fluorescent staining.
8. We use GenePix Pro 7 software for the detection and quantification of each individual spot.
9. Antibody performance such as specificity, functionality, and stability should be validated using western blot prior to array experiments.
10. The spotting concentration should be optimized for each antibody and surface chemistry of the slide. The working concentrations of capture antibodies are generally ranging from 2 to 20 μg/ml.
11. The printing layout should include target antibodies, positive and negative controls with duplicates.
12. Hold the slide edges only, and avoid touching the slide surface. Handle and dry the slides in a clean environment.
13. The storage time for printed microarrays could be determined by the stability of the capture antibodies. The arrays could be directly detected within 2 h after printing, or stored in 4 °C or −20 °C in dark prior to use.
14. The sealed 384-well sample plates can be stored at 4 °C for short-period storage or −80 °C for long-time storage.
15. Avoid drying of the array slide between blocking, incubation, or washing steps, otherwise it will cause high background noises.
16. For biotin-labeled samples, caution must be taken to circumvent any contaminations by amines or sodium azide. An optimal molar ratio of biotin: protein is the key for the success of this experiment to ensure the maximum antigen-antibody interactions for best results.
17. The optimal dilution of serum samples relies on the concentration of the target protein, types of labeling reagent, slide type, or the composition of blocking buffer. Signal-to-noise ratios will have to be optimized for each experiment. A concentration of 0.2 mg/ml biotin in serum seems a workable strategy in general for direct labeling of serum samples.
18. From this step on, wrap the chambers with aluminum foil to avoid array exposure to light.

19. The laser and PMT gain settings could be optimized so that the signal is sufficiently bright, but not oversaturated. This procedure could be adjusted based on the preview results.
20. Modification of image brightness or contrast does not have any effect on the final intensity value; it is only used to enhance visibility of the spots on the screen.
21. Normalization is critical for the data analysis so that a comparison of data generated from different sub-arrays, slides, and dyes could be reasonably performed.
22. The Super Curve software can be obtained at http://bioinformatics.mdanderson.org/main/OOMPA:Overview.
23. SNR = (Average signal intensity − Average background intensity)/Standard deviations of background signals [6].

Acknowledgments

This work was partly supported by a grant from the Lupus Research Institute to T.W. and a startup fund from the University of Houston to T.W.

References

1. Orchekowski R, Hamelinck D, Li L, Gliwa E, VanBrocklin M, Marrero JA, Woude GFV, Feng Z, Brand R, Haab BB (2005) Antibody microarray profiling reveals individual and combined serum proteins associated with pancreatic cancer. Cancer Res 65(23):11193–11202
2. Borrebaeck CA, Wingren C (2014) Antibody array generation and use. Methods Mol Biol 1131:563–571
3. Mueller C, Liotta LA, Espina V (2010) Reverse phase protein microarrays advance to use in clinical trials. Mol Oncol 4(6):461–481
4. Wu T, Ding H, Han J, Arriens C, Wei C, Han W, Pedroza C, Jiang S, Anolik J, Petri M, Sanz I, Saxena R, Mohan C (2016) Antibody-array-based proteomic screening of serum markers in systemic lupus erythematosus: a discovery study. J Proteome Res 15(7):2102–2114. doi:10.1021/acs.jproteome.5b00905
5. Wu T, Du Y, Han J, Singh S, Xie C, Guo Y, Zhou XJ, Ahn C, Saxena R, Mohan C (2013) Urinary angiostatin—a novel putative marker of renal pathology chronicity in lupus nephritis. Mol Cell Proteomics 12(5):1170–1179. doi:10.1074/mcp.M112.021667
6. Zong Y, Zhang S, Chen H-T, Zong Y, Shi Y (2007) Forward-phase and reverse-phase protein microarray. Methods Mol Biol 381:363–374

Chapter 20

Protein Arrays II: Antigen Arrays

Yulin Yuan, Hongting Wang, Zuan-Tao Lin, Xia Hong, Mikala Heon, and Tianfu Wu

Abstract

Antigen arrays are fabricated using various antigens such as DNA, histones, synthetic peptides, recombinant proteins, or cell extracts to detect autoantibodies in autoimmune diseases, alloantibodies in transplantation, drug-induced antibodies or cancer-induced antibodies in blood or cell culture supernatant. In this protocol, we will provide a step-by-step executable procedure to perform antigen arrays, including antigen preparation and printing, blocking, sample loading, array detection, imaging, and data analysis.

Key words Antigen arrays, Autoantibody profiling, Biomarker

1 Introduction

Various antigens can be printed onto microarray slides to react with corresponding antibodies in serum samples. Reactivity patterns are implacable in guiding the discovery of novel antigens or the identification of antigenic epitopes, allergens, or vaccine targets [1], or profiling of autoantibodies [2].

There are two major strategies to fabricate protein antigen arrays: recombinant protein-based or cDNA-based arrays. The latter is also called Nucleic Acid Programmable Protein Array (NAPPA) which was well described elsewhere [3]. In this chapter, we will only describe the step-by-step procedure of the development and experimental application of antigen arrays based on recombinant proteins or other antigens as illustrated in Fig. 1.

The integration of plasmonic surfaces into antigen arrays has shown promising results in improving the sensitivity and robustness of detection [4, 5]. Currently plasmonic slides are commercially available in Plasmonix, Inc. (Gaithersburg, MD), and this may be helpful in enabling the paradigm shift of the array detection if confirmed.

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_20,

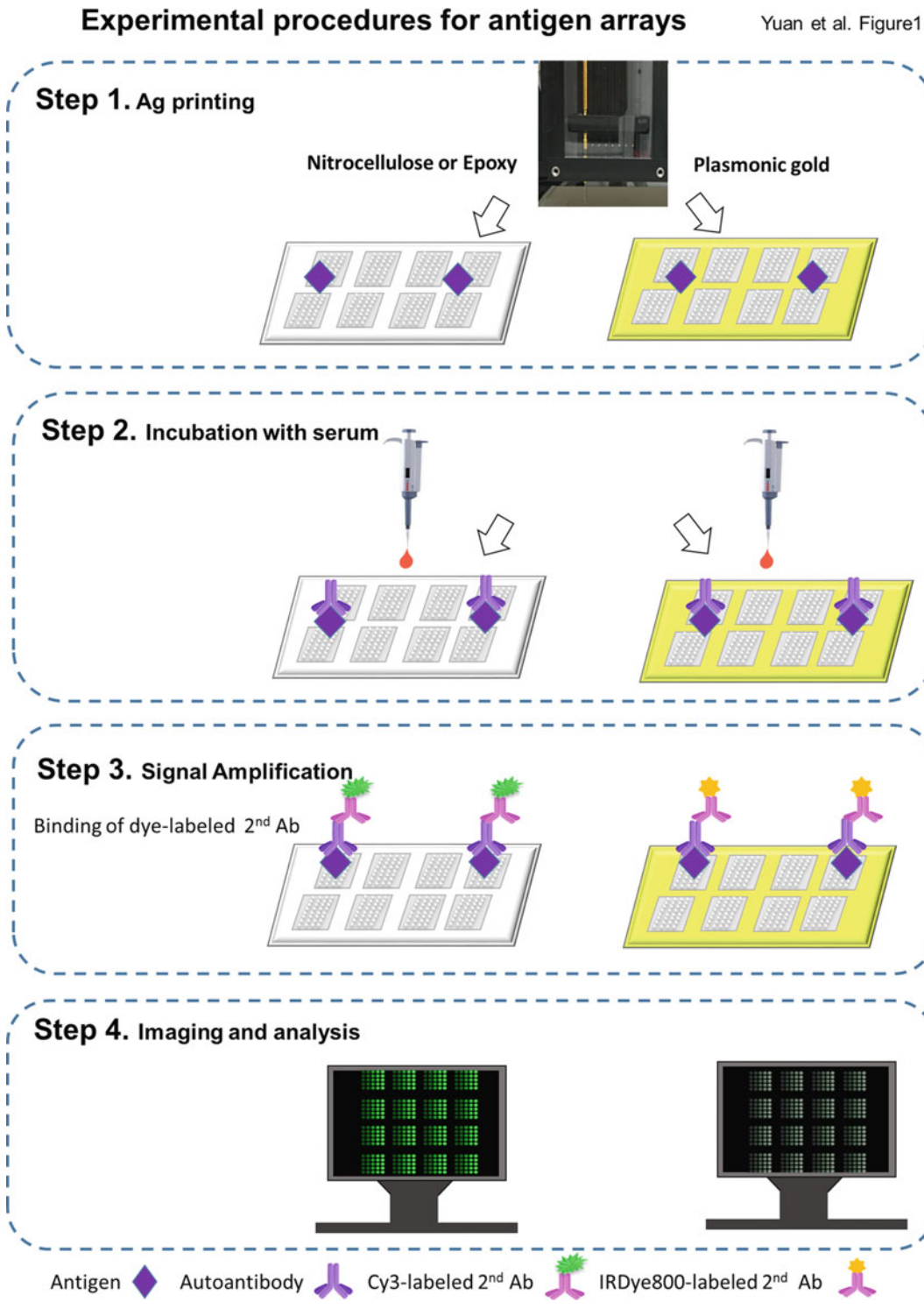


Fig. 1 Workflow of Antigen arrays. Antigens from a recombinant or non-recombinant source can be printed onto nitrocellulose or epoxy-coated glass slides, plastic slides, or plasmonic gold chips, followed by the incubation with serum samples and the detection with fluorescent dye-labeled secondary antibody. The plasmonic gold surfaces are expected to result in enhanced fluorescence signals. *Ab* antibody, *Ag* antigen

2 Materials

Milli-Q water was used throughout.

PBS: 137 mM NaCl, 2.7 mM KCl, 4.3 mM Na_2HPO_4, 1.47 mM KH_2PO_4.

PBST: 137 mM NaCl, 2.7 mM KCl, 4.3 mM Na_2HPO_4, 1.47 mM KH_2PO_4, 0.1% Tween-20, pH 7.4.

2.1 Antigen Preparation

1. Antigens can be proteins, peptides, carbohydrates, nucleic acids, or even pathogen lysates (*see* **Note 1**).
2. Antigen dilution buffer: PBS.

2.2 Antigen Printing

1. Printing buffer: PBS.
2. Slide: Sony DADC Epoxy MS slide (STRATEC Consumables GmbH, Austria).
3. SciFlexarrayer S3 (Scienion, Berlin, Germany).
4. 384-Well microarray microplates (Arrayit Corporation, USA).
5. Positive control: Human IgG purified from human serum.
6. Negative control: printing buffer.

2.3 Sample Preparation

1. Washing buffer: PBST.
2. Serum diluent buffer: the same as blocking buffer.
3. Serum samples (*see* **Note 2**).

2.4 Array Detection

1. Washing buffer: PBST.
2. Blocking buffer: 1% BSA in PBST or 3% nonfat milk in PBST, or super G blocking buffer (Grace bio-labs, Inc., USA).
3. Detection antibodies: Anti-Human IgG (Fc specific)-Cy3 antibody (Jackson ImmunoResearch Laboratories, Inc., USA) (*see* **Note 3**).

2.5 Scanning and Analysis

1. Scanner: GenePix Microarray Scanner 4400A (Molecular Devices LLC, USA).
2. Analysis software: GenePix Pro 7 software (Molecular Devices LLC, USA).

3 Method

3.1 Antigen Preparation

1. Reconstitute the antigens in PBS.
2. Serially dilute the antigens and controls with printing buffer to a suitable concentration (*see* **Note 4**).

3.2 Printing of Antigens

1. Transfer the antigens into a 384-well microplate with 40 μl per well.
2. Cover the slide with Microseal® 'B' Adhesive Seals and centrifuge the microplate 2000 × *g* for 5 min at 4 °C to remove any bubbles.
3. Fill the wash bottle of SciFlexarrayer S3 with millipore water and empty the waste bottle.
4. Set the relative humidity to 50% and ambient temperature to 4 °C in the printing chamber.
5. Turn on the SciFlexarrayer S3 and open the software.
6. Choose the "prime" option to initiate a standard starting routine.
7. Click "Do Task," then "sciCLEANWash Tray" to wash the Piezo Dispense Capillary (PDC).
8. Mount Epoxy slides onto the SciFlexarrayer S3.
9. Target setup
   (a) Target: Select the fields where antibodies will be printed.
   (b) Field setup: Based on the total number of antibodies to be printed to set the work table; for each antibody, set a twofold dilution gradient.
10. Click "Run" to start printing.
11. When printing is finished, wash the PDC three times with 1% sciCLEAN 8.
12. Use the slides after air drying for 20 min, or store them at 4 °C.
13. Cover the 384-well microarray microplates with Microseal® 'B' Adhesive Seals and store the plates in −80 °C.

3.3 Blocking (See Note 5)

1. Assemble the slides into the 16 Well ProPlate® Slide Modules (Grace Bio-Labs, Inc.).
2. Add 100 μl blocking buffer to each sub-array.
3. Cover the incubation chamber with ProPlate® Slide Module Seal Strips and incubate the slides at room temperature for 60 min with gentle shaking (55 rpm) (*see* **Note 6**).
4. Aspirate blocking buffer from each well (*see* **Note 7**).
5. Add 100 μl/well washing buffer and incubate on an orbital shaker rotating at room temperature for 5 min (55 rpm) (*see* **Note 8**).

3.4 Array Detection

1. Thaw serum samples and keep them on ice until use.
2. Dilute serum sample 1:100 in serum diluent buffer (*see* **Note 9**).
3. Aspirate washing buffer from each well.

4. Add 75 μl diluted serum samples into each well and cover the incubation chamber with adhesive film.
5. Incubate the slides in a humidified chamber with shaking (55 rpm) overnight at 4 °C (*see* **Note 10**).
6. Add 100 μl/well washing buffer and incubate on an orbital shaker rotating at room temperature for 5 min (55 rpm), repeat this wash step four times.
7. Dilute Cy3-conjugated anti-human IgG antibody 5000-fold with blocking buffer.
8. Add 100 μl diluted antibody to each well after removing washing buffer.
9. Cover slides with ProPlate® Slide Module Seal Strips and keep it in a dark chamber.
10. Incubate the arrays at room temperature for 60 min with gentle shaking (55 rpm).
11. Wash the arrays using 100 μl/well wash buffer (5 × 5 min) with shaking (55 rpm).
12. Wash the arrays using 100 μl/well PBS (3 × 5 min) with shaking (55 rpm).
13. Wash the arrays using 100 μl/well H_2O for 5 min with shaking (55 rpm).
14. Remove slides from ProPlate® 16 Well Slide Module and immerse the each slide in a separate 50 ml conical tube with washing buffer, shake at room temperature for 5 min (55 rpm).
15. Discard the washing buffer and add 45 ml H_2O, shake at room temperature for 5 min (55 rpm).
16. Discard the liquid and centrifuge the conical tube at 2000 × *g* for 2 min to remove liquids on the slide (*see* **Note 11**).
17. Place the slides into a clean slide holder to dry in air or under a stream of nitrogen gas.
18. Scan or store the slides at room temperature in a dark chamber.

3.5 Scan and Analysis

The scan and analysis procedures are the same as described in Subheading 3.7 in Chapter 19.

4 Notes

1. All antigens are prepared with antigen diluent and filtered with a 0.22 μm filter.
2. The serum diluent buffer also needs to be filtered with a 0.45 μm filter to remove undissolved particles or aggregates that can cause increased and uneven background or noise [6].

3. Depending on the purpose of detection, other fluorescently labeled secondary antibodies, such as anti-IgA, IgD, IgG, IgM, or IgG subclasses (IgG1, IgG2, IgG3, or IgG4), can be used. Various Ig classes can also be detected by using different fluorescent dyes, but the absorption or emission spectra for detection should not overlap. GenePix Microarray Scanner 4400A is equipped with four-channel lasers which allows for a simultaneous detection of four different classes of Immunoglobulins or subclasses of IgG if the detection antibodies are labeled with four non-overlapped fluorescent colors.
4. Antigens should be printed at least in duplicate, along with antigen diluent as negative controls. Our spotting concentration of antigens ranges from 6.25 to 100 μg/ml.
5. Blocking prior to the addition of serum samples can reduce overall background signals.
6. Place the slides in a humid chamber filled with wet filter paper to avoid evaporation of reagents during incubation.
7. Avoid touching the printed area of the array with a pipette, only touch the corners of each chamber.
8. Do not allow the slide to completely dry out. During incubation avoid foaming and remove any bubbles by centrifugation and a pipette.
9. Human serum samples might be infectious; hence, biohazard guidelines should be followed regarding waste disposal, treatment, and the disinfection of reusable materials.
10. Sample incubation overnight is helpful in increasing antigen-antibody binding to improve signal intensity.
11. Spinning the slide in low speed will help its drying. Alternatively, slides can be dried in a safety cabinet with the airflow on.

Acknowledgments

This work was partly supported by a grant from the Lupus Research Institute to T.W. and a startup fund from the University of Houston to T.W.

References

1. Prechl J, Papp K, Erdei A (2010) Antigen microarrays: descriptive chemistry or functional immunomics? Trends Immunol 31(4):133–137
2. Li QZ, Xie C, Wu T, Mackay M, Aranow C, Putterman C, Mohan C (2005) Identification of autoantibody clusters that best predict lupus disease activity using glomerular proteome arrays. J Clin Invest 115(12):3428–3439. doi:10.1172/JCI23587
3. Ramachandran N, Raphael JV, Hainsworth E, Demirkan G, Fuentes MG, Rolfs A, Hu Y, LaBaer J (2008) Next-generation high-density self-assembling functional protein arrays. Nat Methods 5(6):535–538

4. Tabakman SM, Lau L, Robinson JT, Price J, Sherlock SP, Wang H, Zhang B, Chen Z, Tangsombatvisit S, Jarrell JA, Utz PJ, Dai H (2011) Plasmonic substrates for multiplexed protein microarrays with femtomolar sensitivity and broad dynamic range. Nat Commun 2:466. doi:10.1038/ncomms1477
5. Zhang B, Kumar RB, Dai H, Feldman BJ (2014) A plasmonic chip for biomarker discovery and diagnosis of type 1 diabetes. Nat Med 20 (8):948–953. doi:10.1038/nm.3619
6. Papp K, Prechl J (2012) The use of antigen microarrays in antibody profiling. Methods Mol Biol 815:175–185

Chapter 21

Protein Arrays III: Reverse-Phase Protein Arrays

Yulin Yuan, Xia Hong, Zuan-Tao Lin, Hongting Wang, Mikala Heon, and Tianfu Wu

Abstract

The reverse-phase protein array (RPPA) is to use highly specific antibodies to interrogate pan or post-translationally modified protein targets, such as phosphorylated proteins, particularly the proteins involved in cell signaling pathways. In this protocol we will cover the preparation of cell (or tissue) lysates, sample printing, antibody validation, antibody interrogation, signal amplification steps, imaging and data analysis. In this protocol, colorimetric catalyzed signal amplification (CSA) chemistry, fluorescence and near-infrared (NIR) based detection methods will be described.

Key words Reverse-phase protein array, Antibody validation, Cell signaling pathways, Near-infrared (NIR) fluorescence

1 Introduction

RPPA is a high-throughput antibody-based technique to detect protein expression in cell or tissue lysates, similar to Western blots [1]. Western blot has been widely used historically for the detection of the expression of single proteins; however, the need of a relatively large amount of protein samples per run makes this method unsuitable for precious and limited clinical samples. Therefore, there is a great need in improving the sensitivity of detection strategy. In addition, to maximize the use of the precious clinical samples, a multiplex assay must be developed. The design of RPPA technologies allows for an increased sensitivity, minimal requirement of samples, and multiplexity of the assay. Recent studies have shown that RPPA is promising in the application of ultrasensitive detection of critical proteins or markers in biological samples or clinical samples [2–7]. The advantages of RPPA include the possibility of personalized molecular profiling for patients using automated, high-throughput robotic arraying system, minimal amount of clinical specimens, and high sensitivity. This technology allows for the detection of protein samples extracted from limited blood cells

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_21,

from patients, or laser capture micro-dissected biopsies, cell culture, serum, urine CSF, synovial fluid, and vitreous humor. Depending on the type of the arrayer system, down to 20 pg ~ 1 ng of protein samples could be deposited, and several thousand samples can be analyzed simultaneously on the same slide [1].

There are various detection methods available [8], and the most popular technologies include colorimetric such as catalyzed signal amplification (CSA) chemistry (DAKO) fluorescence and near-infrared (NIR) methods as illustrated in Fig. 1. An obvious advantage of colorimetric methods lies in its simplicity of spot imaging, where a regular flatbed scanner is sufficient. Fluorescence detection is advantageous in terms of the commercial availability of various fluorescent dyes and the great brightness and high sensitivity [9]. The NIR detection provides the largest dynamic range (up to 4 orders of magnitudes) of signal-to-noise ratio.

2 Materials

Milli-Q water was used throughout.

PBS: 137 mM NaCl, 2.7 mM KCl, 4.3 mM Na_2HPO_4, 1.47 mM KH_2PO_4.

PBST: 137 mM NaCl, 2.7 mM KCl, 4.3 mM Na_2HPO_4, 1.47 mM KH_2PO_4, 0.1% Tween-20, pH 7.4.

2.1 Sample Preparation

Various cell lysates or tissue lysate can be printed onto slides for RPPA.

1. RIPA buffer (Cell Signaling Technology, Inc.) for tissue lysates and cell lysis buffer (Cell Signaling Technology, Inc.) for cell lysates.
2. Halt™ Protease and Phosphatase Inhibitor Cocktail, EDTA-free (100×), (Thermo Fisher Scientific Inc., USA)
3. 50 μl 2-mercaptoethanol (final concentration 2.5% v/v).
4. Novex® Tris-Glycine SDS Sample Buffer (2×) (Thermo Fisher Scientific Inc., USA).
5. Pierce™ BCA Protein Assay Kit (Thermo Fisher Scientific Inc., USA).

2.2 Printing

1. Print buffer: PBS.
2. Slide: Sony DADC Epoxy MS slide (STRATEC Consumables GmbH, Austria).
3. SciFlexarrayer S3 (Scienion, Berlin, Germany).

2.3 Immunostaining

1. Antigen solution buffer: PBS.
2. Blocking solution: 1% BSA in PBST or 3% nonfat milk in PBST, or super G blocking buffer (Grace bio-labs, Inc., USA).

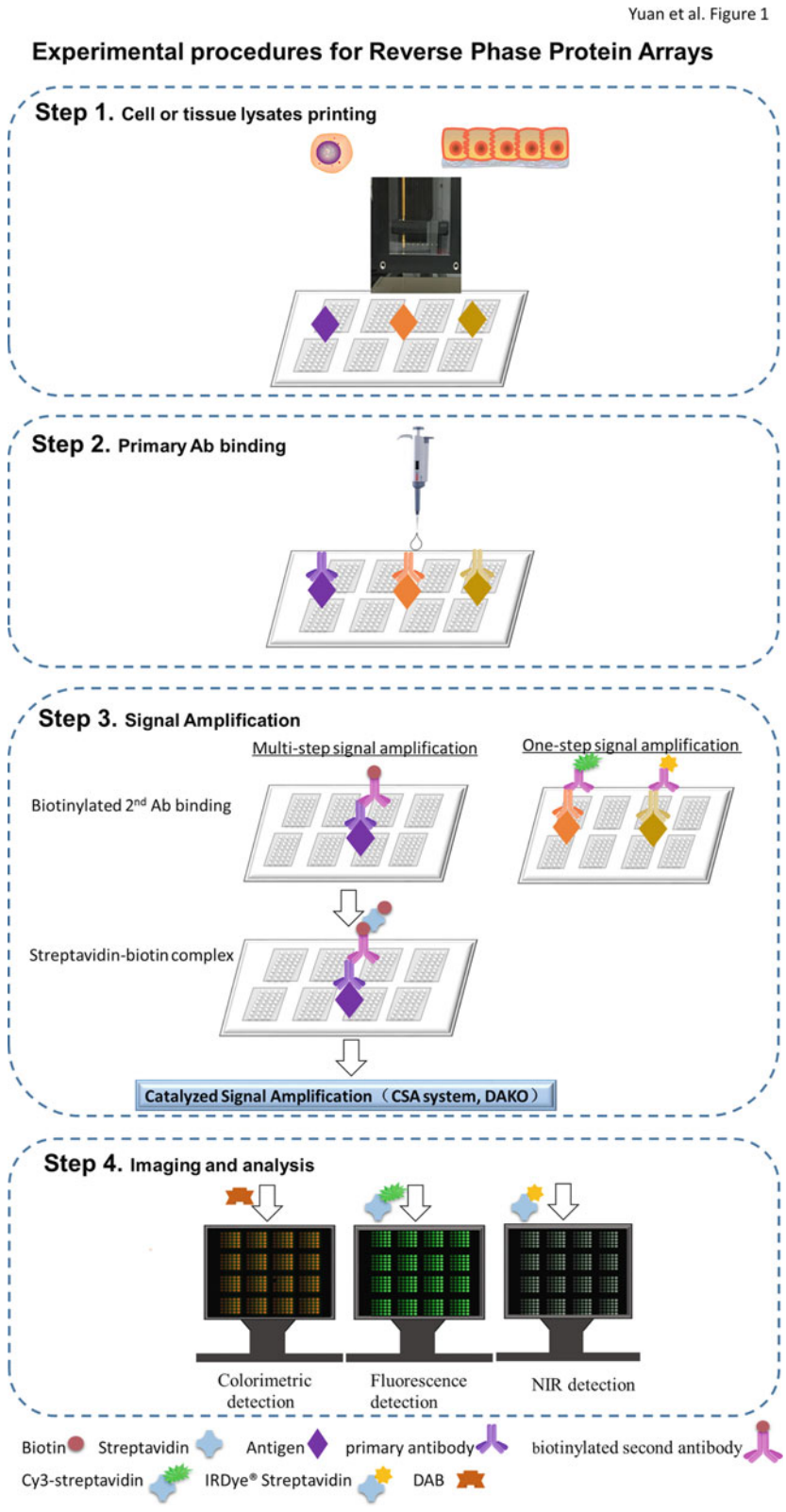


Fig. 1 Workflow of reverse-phase protein arrays (RPPA). Crude extracts from cell lines or tissues are arrayed on glass slides. Upon specific primary antibody binding, the array could be detected either via traditional one-

3. Validated primary antibodies (*see* **Note 1**).
4. Biotinylated secondary antibody.
5. Biotin Blocking System (Dako, USA): 0.1% Avidin, 1% Biotin.
6. CSA, Catalyzed Signal Amplification System (Dako, USA) (*see* Subheading 4): Peroxidase Block, Protein Block, Link Antibody, Streptavidin-Biotin Complex, Amplification Reagent, Streptavidin-Peroxidase and substrate DAB chromogen.
7. Streptavidin (Cy3) (GeneTex, Inc., USA) or IRDye® 800CW (LI-COR, Inc., USA).
8. Loading control: purified proteins or recombinant peptides (such as β-actin).
9. Negative control: cell lysis buffer.

2.4 Quantification of Total Proteins

1. SYPRO® Ruby Protein Blot Stain (Thermo Fisher Scientific Inc., USA).
2. Fixative solution: 41.5 ml H_2O, 3.5 ml acetic acid (final concentration 7% v/v), 5 ml methanol (final concentration 10% v/v). Store tightly closed at room temperature.
3. Fluorescent image capturing system, such as a UV transilluminator or laser scanner (excitation 280 nm, emission 618 nm).

2.5 Imaging and Data Analysis

1. According to different detection methods, optical flatbed scanner for colorimetric detections, GenePix Microarray Scanner 4400A (Molecular Devices LLC, USA) for fluorescence detection or InnoScan 710-IR scanner (Innopsys) for near-infrared detection.
2. Analysis software: GenePix Pro 7 software (Molecular Devices LLC, USA).

3 Methods

3.1 Protein Extraction

3.1.1 Cell Lysis

1. Harvest cells into a 1.5 ml tube. Wash cells three times with cold PBS (*see* **Note 2**).
2. Spin the pellet at 4 °C for 10 min at 1200 × *g* to remove excess buffer.
3. Add 100 μl cell lysis buffers per 1×10^6 cells (cell lysis buffer containing Halt™ Protease and Phosphatase Inhibitor Cocktail, Roche Ltd., Switzerland).

Fig. 1 (continued) step signal amplification using dye-labeled secondary antibody or via a more sophisticated multiple signal amplification using a tyramide-based CSA system (DAKO). This CSA system allows for significant amplification of signals from colorimetric substrate DAB, fluorescence Cy3 or NIR IRDye® 800CW dyes. *Ab* antibody. *NIR* near infrared

4. Vortex the tube for 30 s and store it on ice for 30 min.
5. Spin at 15,000 × *g* for 20 min at 4 °C.
6. Collect the supernatant fraction for immediate use or aliquot it and freeze them away.

3.1.2 Sample Preparation

1. Determine the total protein concentration with a Pierce™ BCA Protein Assay Kit.
2. Adjust the total protein concentration of the lysates to 0.5–1 mg/ml.
3. Mix the lysates and 2× Novex® Tris-Glycine SDS Sample Buffer (1:1).
4. Heat the mixture in a dry heat block at 100 °C for 5 min.
5. Centrifuge the mixture for 1 min at 2000 × *g* at room temperature.
6. Use these lysates to print arrays or store them at −80 °C until use.

3.2 Lysate Printing

All samples are printed in a dilution curve on the array to match the protein concentration with the antibody affinity (*see* **Note 3**).

1. Transfer the diluted lysates into 384-well plates.
2. Cover the slide with Microseal® 'B' Adhesive Seals and centrifuge the microplate 2000 × *g* for 5 min at 4 °C to remove any bubbles.
3. Fill SciFlexarrayer S3 wash bottle with fresh milliQ water and empty the waste bottle.
4. Set the relative humidity to 50% and ambient temperature to 4 °C in the printing chamber.
5. Turn on the SciFlexarrayer S3 and open the software.
6. Choose the "prime" option to initiate a standard starting routine.
7. Click "Do Task," then "sciCLEANWash Tray" to wash the Piezo Dispense Capillary (PDC).
8. Mount Epoxy slides onto the SciFlexarrayer S3 slide stage.
9. Target setup
   (a) Target: Select the fields where lysates will be printed.
   (b) Field setup: According to the number of lysate samples to set the work table; for each sample, set a twofold dilution gradient.
10. Click "Run" to start printing.
11. When printing is finished, wash the PDC three times with 1% sciCLEAN 8.
12. Use the slides after air-dry for 20 min, or store them at 4 °C.

13. Cover the 384-well microarray microplates with Microseal® 'B' Adhesive Seals and store the plates in −80 °C.

3.3 Determination of Total Protein Concentration on the Slide: Sypro Ruby Staining

1. Wash the slides with H_2O for 5 min with shaking (55 rpm).
2. Immerse the entire slide in a new 50 ml tube containing fixing buffer, shake at room temperature for 15 min (55 rpm).
3. Discard the fixing buffer and wash the slides with H_2O for 5 min with shaking. Repeat this step four times.
4. Incubate the slides in a dark box with Sypro Ruby staining solution for 30 min.
5. Rinse the slides with H_2O to remove excessive dye.
6. Keep the slides in a dark box to dry.
7. Scan the slides using a fluorescence imaging system.

3.4 Antibody Validation

The reliability of RPPA results largely depends on the quality of the antibodies used. Therefore, all antibodies should be validated with Western blot first. Antibodies suitable for RPPA must show a single master band. So far, several hundred commercial antibodies have been validated for the purpose of RPPA [2, 4, 10, 11] (*see* **Note 4**).

1. Cells with or without stimulation (agonist or antagonist of a particular protein) are lysed and run on a 4–15% gradient SDS-PAGE.
2. Proteins are transferred onto a PVDF membrane using a Semi-Dry and Rapid Bloting system (Bio-rad Laboratories Inc., USA).
3. Block for 1 h at room temperature.
4. Incubate the membrane in primary antibodies overnight at 4 °C with shaking.
5. Wash membrane with PBST at room temperature 3× for 10 min.
6. Dilute fluorescently labeled secondary antibodies to 1:5000~10,000 with blocking buffer.
7. Incubate membrane with diluted secondary antibodies at room temperature for 45 min with gentle shaking. Keep the membrane in the dark from this step forward.
8. Wash membrane with PBST at room temperature 3× for 5 min.
9. Get the membrane ready, and scan it (*see* **Note 5**).
10. The antibodies which are able to generate a single predominant band at the correct molecular weight will be selected to perform RPPA.

3.5 Immunostaining

3.5.1 RPPA Slide Blocking (See Note 6)

1. Take out the slides from the freezer or refrigerator and place them at room temperature for 10 min.
2. Attach the slides onto a 16 well slide module.

   For CSA-based signal amplification, the blocking steps for biotin, avidin, and endogenous peroxidase are included according to the manufacturer's instructions (*see* **Note 7**).
3. Add 100 µl protein block buffer to each well and incubate at room temperature for 60 min on a shaker (55 rpm).
4. Wash the slides using 100 µl/well washing buffer (3 × 5 min) with shaking (55 rpm)

3.5.2 Array Detection

1. Incubate the slides with 75 µl/well primary antibody or positive/negative control reagent in a humid chamber at 4 °C overnight with shaking (55 rpm).
2. Wash the slides using 100 µl/well washing buffer (3 × 5 min) with shaking (55 rpm).

   Fluorescence/near-infrared detection
   Alternative A: One-step signal amplification
3. Add 50 µl/well Cy3 or IRDye® 800CW labeled second antibody and incubate for 60 min with shaking (55 rpm).
4. Wash the arrays using 100 µl/well wash buffer (5 × 5 min) with shaking (55 rpm).
5. Wash the arrays using 100 µl/well PBS (3 × 5 min) with shaking (55 rpm).
6. Wash the arrays using 100 µl/well H_2O for 5 min with shaking (55 rpm).
7. Remove slides from ProPlate® 16 Well Slide Module and immerse the entire slides into washing buffer in a separate 50 ml conical tube, shake at room temperature for 5 min (55 rpm).
8. Discard the washing buffer and add 45 ml H_2O, shake at room temperature for 5 min (55 rpm).
9. Discard the liquid and centrifuge the conical tube at 2000 × *g* for 2 min.
10. Place the slides into a new conical tube to dry in air or under a stream of nitrogen gas.
11. Scan or store the slide at room temperature in a dark chamber.

    Alternative B: Multi-step signal amplification
3. Incubate slides with 75 µl/well biotinylated secondary antibody.
4. Repeat **step 4** of Subheading 3.5.1.

5. Add 50 μl/well streptavidin-biotin complex and incubate for 15 min with shaking (55 rpm) (*see* **Note 8**).
6. Repeat **step 4** of Subheading 3.5.1.
7. Add 50 μl/well Streptavidin-HRP to incubate for 15 min with shaking (55 rpm).
8. Repeat **step 4** of Subheading 3.5.1.
9. Add 50 μl/well amplification reagent (biotinyl-tyramide) and incubate for 15 min.
10. Repeat **step 4** of Subheading 3.5.1.
11. Add 50 μl/well streptavidins (Cy3) or IRDye® Streptavidins and incubate for 15 min with shaking (55 rpm).
12. Repeat **steps 4–11** of one-step signal amplification.

Alternative C: Colorimetric method

3. Incubate slides with 75 μl/well biotinylated secondary antibody.
4. Repeat **step 4** of Subheading 3.5.1.
5. Add 50 μl/well streptavidin-biotin complex and incubate for 15 min with shaking (55 rpm).
6. Repeat **step 4** of Subheading 3.5.1.
7. Add 50 μl/well Streptavidin-HRP to incubate for 15 min with shaking (55 rpm).
8. Repeat **step 4** of Subheading 3.5.1.
9. Add 50 μl/well amplification reagent (biotinyl-tyramide) and incubate for 15 min.
10. Repeat **step 4** of Subheading 3.5.1.
11. Add 50 μl/well DAB substrate-chromogen solution and incubate for 10 min with shaking (*see* **Note 9**).
12. Wash the arrays using 100 μl/well H_2O for 5 min with shaking (55 rpm).
13. Remove the slides from the frame assemblies, and immerse the entire slides in a hematoxylin tube.
14. Rinse gently in a H_2O bath.
15. Dip slides ten times into 0.037 mol/l ammonia solution.
16. Rinse slides in H_2O for 5 min.
17. Discard the liquid and centrifuge the tube 2000 × *g* for 2 min.
18. Put the slides into a new tube to dry in air or under a stream of nitrogen gas.
19. Store the slides in the dark at RT prior to scanning with a flatbed HPScanJet scanner with 600 dpi resolution.

3.6 Imaging and Analysis

To image fluorescently labeled arrays, following steps are performed using GenePix 4400A scanner as an example:

1. Turn on the GenePix 4400A scanner and warm up for 20 min.
2. Insert dry slides upside down on the holder of the scanner.
3. Perform a preview scan of the entire slide.
4. According to the preview results of overall signal intensity and resolution, laser power and PMT gain will be optimized.
5. Scan the slide with the selected settings and save the image.
6. Open the images of the microarrays using GenePix Pro 7 software and load up the array list (GAL file).
7. Adjust the brightness and contrast.
8. Normalize the dataset of signal intensity throughout the slides using positive controls.
9. Assess the sensitivity of the antibody and the linearity of the detected signal by a dilution series of each sample to build a dilution curve.
10. Calculate the average of the duplicate spots to obtain a value for the array. Subtract the local background and calculate the signal-to-noise ratio (SNR).
11. The signal intensity values represent the protein expression level.

4 Notes

1. Primary antibodies must be validated prior to a RPPA experiment. Several validated antibody lists have been published as supplementary information [2, 4, 10, 11], where antibody name, vendor, catalogue number, and clone information were included so that one can directly refer to these information and tremendously save time and efforts in identifying antibody sources.
2. Cultured cells should be pelleted and washed with PBS to remove immunoglobulins, serum, or other contaminating reagents in the medium.
3. The protein cell lysates usually are serially diluted: 1:2, 1:4, 1:8, 1:16, 1:32, 1:64, and 1:128 [2].
4. A validated antibody list is published online on the website of the Protein Microarray Core facility at the University of Texas M.D. Anderson Cancer Center: https://www.mdanderson.org/research/research-resources/core-facilities/functional-proteomics-rppa-core/antibody-information-and-protocols.html.

5. Allowing the membrane to dry may be helpful for enhancing the signal and reducing background noise, but this is not recommended for the purpose of stripping and re-probing.
6. The printed RPPA slides should be blocked prior to the staining procedure.
7. In the conventional biotin-avidin techniques, the evaluation of specific staining can be impeded by the presence of endogenous biotin. Therefore, the Biotin Blocking System is recommended to reduce nonspecific staining due to the endogenous biotin binding activity.
8. The Streptavidin-Biotin Complex should be prepared at least 30 min prior to use.
9. DAB waste will be collected in a hazardous materials container for proper disposal.

Acknowledgments

This work was partly supported by a grant from the Lupus Research Institute to T.W. and a startup fund from the University of Houston to T.W.

References

1. Pierobon M, VanMeter AJ, Moroni N, Galdi F, Petricoin EF (2012) Reverse-phase protein microarrays. Methods Mol Biol 823:215–235
2. Tibes R, Qiu Y, Lu Y, Hennessy B, Andreeff M, Mills GB, Kornblau SM (2006) Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Mol Cancer Ther 5(10):2512–2521. doi:10.1158/1535-7163.MCT-06-0334
3. Sheehan KM, Calvert VS, Kay EW, Lu Y, Fishman D, Espina V, Aquino J, Speer R, Araujo R, Mills GB, Liotta LA, Petricoin EF 3rd, Wulfkuhle JD (2005) Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma. Mol Cell Proteomics 4(4):346–355. doi:10.1074/mcp.T500003-MCP200
4. Kornblau SM, Tibes R, Qiu YH, Chen W, Kantarjian HM, Andreeff M, Coombes KR, Mills GB (2009) Functional proteomic profiling of AML predicts response and survival. Blood 113(1):154–164. doi:10.1182/blood-2007-10-119438
5. Carter BZ, Qiu Y, Huang X, Diao L, Zhang N, Coombes KR, Mak DH, Konopleva M, Cortes J, Kantarjian HM, Mills GB, Andreeff M, Kornblau SM (2012) Survivin is highly expressed in CD34(+)38(−) leukemic stem/progenitor cells and predicts poor clinical outcomes in AML. Blood 120(1):173–180. doi:10.1182/blood-2012-02-409888
6. Nanos-Webb A, Bui T, Karakas C, Zhang D, Carey JP, Mills GB, Hunt KK, Keyomarsi K (2016) PKCiota promotes ovarian tumor progression through deregulation of cyclin E. Oncogene 35(19):2428–2440. doi:10.1038/onc.2015.301
7. Lui VW, Peyser ND, Ng PK, Hritz J, Zeng Y, Lu Y, Li H, Wang L, Gilbert BR, General IJ, Bahar I, Ju Z, Wang Z, Pendleton KP, Xiao X, Du Y, Vries JK, Hammerman PS, Garraway LA, Mills GB, Johnson DE, Grandis JR (2014) Frequent mutation of receptor protein tyrosine phosphatases provides a mechanism for STAT3 hyperactivation in head and neck cancer. Proc Natl Acad Sci U S A 111(3):1114–1119. doi:10.1073/pnas.1319551111
8. Spurrier B, Ramalingam S, Nishizuka S (2008) Reverse-phase protein lysate microarrays for cell signaling analysis. Nat Protoc 3(11):1796–1808

9. Boellner S, Becker K-F (2015) Reverse phase protein arrays—quantitative assessment of multiple biomarkers in biopsies for clinical use. Microarrays 4(2):98–114
10. Gujral TS, Karp RL, Finski A, Chan M, Schwartz PE, MacBeath G, Sorger P (2013) Profiling phospho-signaling networks in breast cancer using reverse-phase protein arrays. Oncogene 32(29):3470–3476
11. Peng A, Wu T, Zeng C, Rakheja D, Zhu J, Ye T, Hutcheson J, Vaziri ND, Liu Z, Mohan C (2011) Adverse effects of simulated hyper-and hypo-phosphatemia on endothelial cell function and viability. PLoS One 6(8):e23268

Chapter 22

Isolation of Exosomes for the Purpose of Protein Cargo Analysis with the Use of Mass Spectrometry

Monika Pietrowska, Sonja Funk, Marta Gawin, Łukasz Marczak, Agata Abramowicz, Piotr Widłak, and Theresa Whiteside

Abstract

Exosomes are intercellular messengers with a high potential for diagnostic and therapeutic utility. It is believed that exosomes present in body fluids are responsible for providing signals which inhibit immune cells, interfere with antitumor immunity, and thus influence the response to treatment and its effect. One of the most interesting issues in exosome studies is proper addressing of their cargo composed of nucleic acids and proteins. Effective and selective isolation of extracellular vesicles and identification of proteins present in exosomes has turned out to be a challenging aspect of their exploration. Here we propose a novel approach that is based on isolation of exosomes by mini-size-exclusion chromatography which allows efficient, rapid, and reliable isolation of morphologically intact and functionally active exosomes without the need of ultracentrifugation. The purpose of this chapter is to describe a simple and high-throughput method to isolate, purify, and identify exosomal proteins using a mass spectrometry approach. The proposed protocol compiles the expertise of two research groups specialized in exosome research and in mass spectrometry-based proteomics. The protocol combines differential centrifugation followed by ultrafiltration, centrifugation-based filtration, and gel filtration on Sepharose 2B in order to obtain exosomal fractions characterized by only low contamination with albumin.

Key words Albumin removal, Cell culture, Exosomes, Filter-aided sample preparation, Mass spectrometry, Peptide assay, Proteomics, Size exclusion chromatography, Ultrafiltration

1 Introduction

Exosomes are double-layer membrane vesicles having a diameter of several tens of nanometers which are formed in late endosomes otherwise known as multivesicular bodies (MVBs) [1]. There are several lines of evidence that exosomes, released by both tumor and non-tumor cells, may be key players involved in intercellular communication. Moreover, their ability for presentation of antigens and modulation of the immune response as well as possible role in

Monika Pietrowska and Theresa Whiteside contributed equally to this work.

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_22,

development of some neurodegenerative diseases make them a very attractive subject of molecular studies [2]. However, there are still many missing pieces of the puzzle that should be discovered for full understanding of the biological activity and function of exosomes. One of the essential research areas in the exosomes field is characterization of their protein cargo. Previous proteomic studies led to the identification of numerous proteins, either constitutive or occasionally occurring molecules that may be crucial for specific functions of these vesicles. Analysis of exosomes from a wide variety of cells and body fluids has allowed for identification of several functional classes of proteins: membrane adhesion factors, membrane transport/trafficking factors, cytoskeletal components, lysosomal markers, antigen presenting factors, cancer-specific antigens, death receptors, cytokines and cognate receptors, iron transport factors, metabolic enzymes, heat shock proteins, and drug transporters. The presence of specific proteins in tumor-derived exosomes suggests the existence of a protein sorting mechanism during their formation [3]. The presence of specific proteins can reflect the origin of exosomes and their functional role. Proteins present in tumor cell-derived exosomes can be a useful source of cancer biomarkers. It has been shown that exosomes released in vitro from breast carcinoma cells contain HER2, while carcinoembryonic antigen was found in exosomes secreted from colon carcinoma cells. Moreover, MelanA/Mart-1 and gp100 proteins that are expressed in melanoma cells were also found in released exosomes [4]. It was observed that amount and content of exosomes isolated from serum [5], ascites fluids [6], pleural effusions [4], and urine [7] of cancer patients positively correlated with tumor progression. There is no doubt that in-depth characterization of the proteomes of vesicles derived from different types of cancer cells could bring a relevant and timeless knowledge in the field of molecular oncology. However, there are two major challenges in studies focused on exosomal proteomes quantity and purity of the analyte. There are several popular methods of exosome isolation like ultracentrifugation, ultrafiltration, or immuno-capture. All of them have some specific features making them more or less suitable for mass spectrometry applications (reviewed in details in [8]).

Effective preparation and quantification of proteins/peptides for mass spectrometry analysis is an important aspect in processing of biological material. It is crucial for peptide identification yield and for the success of the whole experiment. Unfortunately, a substantial loss of analyte is unavoidable in all processes of sample preparation/purification, which is especially undesirable in the case of low-abundant analytes. At the same time, exact quantification of proteins/peptides intended for mass spectrometry analysis is crucial for the credibility of results, especially when using a label-free strategy. As a consequence, it often means working with trace amounts of biological material, insufficient for the flagship methods

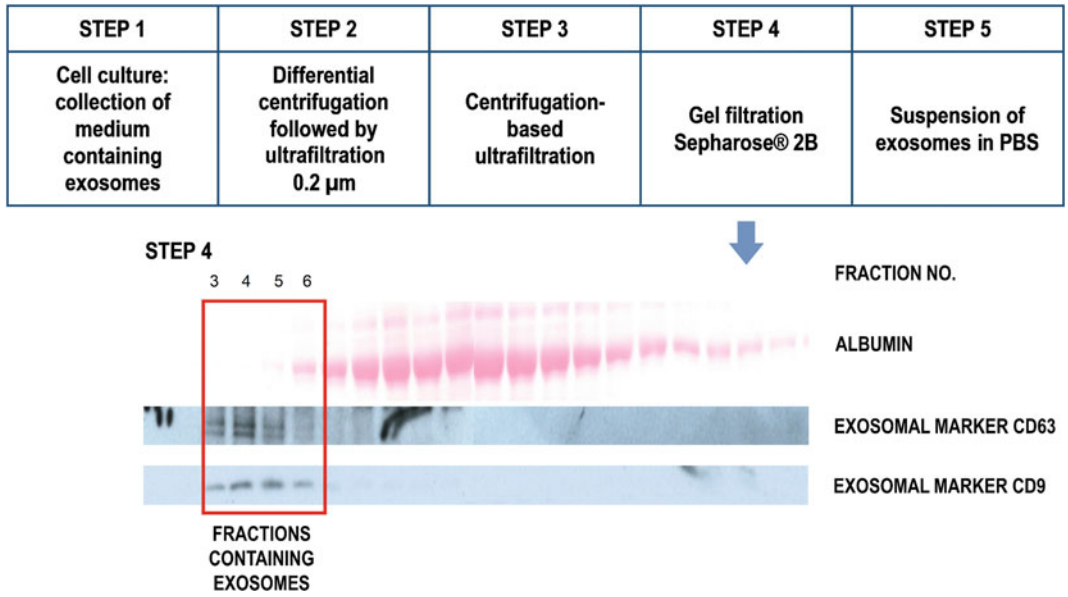


Fig. 1 Preparation of albumin-depleted exosomal fractions

of protein quantification like Bradford or BCA assays. Mass spectrometry is an analytical technique especially suited for analyses of exosomes due to a small amount of available protein material. However, the quality of isolated exosomes is a crucial condition of a successful analysis.

Here we propose a complete and validated protocol for preparation of high quality samples of exosomal proteins for high resolution mass spectrometry analysis. It is a high-throughput method suitable for exosome isolation without the loss of their biological activity (Fig. 1). The pipeline includes well-known steps modified and adapted to low-scale vesicle studies, sensitive protein/peptide quantification by tryptophan fluorescence measurement, protein digestion according to a modified FASP method originally introduced by Wiśniewski et al., and finally qualitative and quantitative analysis by LC-MS/MS resulting in high coverage of a sample proteome.

2 Materials

2.1 Laboratory Equipment

1. Centrifuge equipped with a fixed angle rotor for 1.5/2 ml centrifuge tubes and with adaptors for 50 ml centrifuge tubes, and a swing bucket rotor with adaptors for 250 ml bottles (for Vivacell 100 concentrators).
2. Temperature controlled shaker.
3. Vortex mixer.

4. Vacuum concentrator.
5. Laboratory incubator (working at 37 °C, 5% CO_2).
6. Microplate reader enabling fluorescence excitation in the range of UV light and fluorescence detection in visible light range.
7. LC-MS/MS system:
   (a) MALDI-ToF/ToF MS ultrafleXtreme™ coupled with EASY nLC nano-liquid chromatograph and PROTEINEER fc II fraction collector (all from Bruker Daltonik).
   (b) Hybrid mass spectrometer Q Exactive Quadrupole-Orbitrap coupled with Dionex Ultimate 3000 RSLC nanoLC (both from Thermo Fisher Scientific).

2.2 Isolation of Exosomes for Mass Spectrometry Application

2.2.1 Solutions/Reagents

1. Cell culture medium (e.g., Dulbecco's Modified Eagle's Medium—high glucose, further referred to as DMEM).
2. Fetal bovine serum free from animal exosomes (FBS EXO-) (e.g., Gibco™ Exosome depleted FBS, Thermo Fisher Scientific).
3. Phosphate buffered saline (PBS).
4. Penicillin-Streptomycin (10,000 U/ml).
5. Sepharose® 2B 60–200 μm bead diameter (Sigma-Aldrich).
6. Lysis buffer (exoLB): 6% sodium dodecyl sulfate (SDS), 200 mM dithiothreitol (DTT), 200 mM Tris–HCl pH 7.6.

2.2.2 Consumables

1. 150 cm^2 cell culture flasks, sterile.
2. 50 ml centrifuge tubes.
3. Disposable plastic Pasteur pipettes.
4. 0.22 μm syringe filters with hydrophilic membrane.
5. Vivacell® 100 concentrator units, 100,000 MWCO, PES membrane (Sartorius).
6. Econo-Pac® chromatography columns (Bio-Rad) (*see* **Note 1**).

2.3 Protein/Peptide Assay by Tryptophan Fluorescence Method

1. L-Tryptophan, stock solution: 1 mg/ml in water; working solutions in water: 0.1 mgTrp/ml and 0.01 mgTrp/ml (*see* **Note 2**).
2. 96-Well or 384-well non-treated black flat-bottom polystyrene microplate.

2.4 Protein Digestion and Fractionation

For additional details concerning preparation of solutions, SAX-tip columns, and desalting C18-tip columns for mod-FASP, please refer to [9, 10] and instructions given by the Authors at the webpage of the Max Planck Institute of Biochemistry [11, 12].

2.4.1 Solutions

1. 8 M urea in 0.1 M Tris–HCl pH 8.5.
2. 50 mM iodoacetamide (IAA) in 8 M urea/0.1 M Tris–HCl pH 8.5, or 100 mM IAA in ultrapure water for in-solution digestion.
3. Trypsin (Promega): stock solution 0.5 μg/μl in 50 mM acetic acid for mod-FASP protocol, or 0.1 μg/μl in 50 mM acetic acid for in-solution digestion (*see* **Note 3**).
4. 50 mM Tris–HCl pH 8.5.
5. For in-solution digestion: 100 mM DTT in ultrapure water, and 50 mM NH_4HCO_3 in ultrapure water.
6. Methanol (at least HPLC gradient grade).
7. 1 M NaOH.
8. Britton-Robinson Universal Buffer (BRUB) pH 5 and pH 2 (both diluted five times with water before use).
9. Solutions of trifluoroacetic acid (TFA) in water: 0.1% (v/v), 1% (v/v), 10% (v/v).
10. 60% acetonitrile, 0.1% TFA (v/v) in water.

2.4.2 Consumables and Other Materials

1. Microcon-30 kDa Centrifugal Filter Unit with Ultracel-30 membrane (Merck).
2. Clear polypropylene 200 μl pipette tips, 0.5 and 2 ml reaction tubes (*see* **Note 4**).
3. Scalpel.
4. Empore™ SPE Disks Anion-SR, diam. 47 mm (SUPELCO).
5. Empore™ SPE Disks C18, diam. 47 mm (SUPELCO).
6. Blunt HPLC needle (Hamilton NDL KFga16/51mm/pst3) and a plastic or metal wire.
7. Humid chamber (*see* **Note 5**).

2.5 Mass Spectrometry

1. Highest quality plastic consumables and glass vessels, including test tubes, vials, pipette tips, and bottles for solution storage (*see* **Note 4**).
2. LC-MS grade solvents: water, acetonitrile (ACN).
3. For LC-MALDI MS:
   (a) NS-MP-10 Biosphere C18 pre-column (100 μm × 2 cm, 5 μm granulation, 100 Å) from Nanoseparations (Nieuwkoop, the Netherlands).
   (b) Acclaim PepMap100 C18 analytical column (75 μm × 15 cm, 3 μm granulation, 100 Å, Thermo Scientific).
   (c) MTP AnchorChip 1536 T F 800 μm target plate (Bruker).
   (d) α-Cyano-4-hydroxycinnamic acid (HCCA) matrix for MALDI-TOF MS.

(e) Peptide Calibration Standard II for Mass Spectrometry (Bruker).

(f) 0.05% TFA/H_2O.

(g) 90% ACN, 0.05% TFA.

4. For quadrupole-Orbitrap LC-MS/MS:

(a) Acclaim PepMap100 C18, 5 μm, 100 Å, 300 μm i.d. × 5 mm (Thermo Scientific).

(b) Acclaim PepMap RSLC C18, 2 μm, 100 Å, 75 μm i.d. × 25 cm, nanoViper (Thermo Scientific).

(c) Pierce™ LTQ ESI Positive Ion Calibration Solution (Thermo Fisher Scientific).

(d) 0.1% formic acid/H_2O.

(e) 90% ACN, 0.1% formic acid.

(f) 98% H_2O, 0.1% TFA.

3 Methods

3.1 Isolation of Exosomes for Mass Spectrometry Application

1. Culture HNSCC cell lines in 25 ml DMEM (supplemented with 10% FBS EXO- and 1% Penicillin/Streptomycin) from 30–40% to 70–80% confluency in 150 cm^2 cell culture flasks and collect supernatants after 72 h (*see* **Note 6**).
2. Centrifuge 10 min at 2000 × *g*, room temperature (for dead cells removal).
3. Collect the supernatant carefully (you can use a Pasteur pipette) and transfer it to a fresh centrifuge tube (*see* **Note 7**).
4. Centrifuge 30 min at 10,000 × *g*, 4 °C (for cell debris removal).
5. Collect the supernatant carefully (you can use a Pasteur pipette, you should leave some liquid on the bottom).
6. Filter the collected supernatant using a 0.22 μm syringe filter (for apoptotic bodies and microvesicles removal).
7. Prepare a Vivacell 100 concentrator (*see* **Note 8**).
8. Add 50 ml of the collected filtrate into a Vivacell 100 concentrator unit and centrifuge at 700 × *g*, 35 min, 4 °C, until the concentrate reaches exactly 1 ml.
9. Collect the concentrate containing exosomes from the upper chamber (1 ml).
10. Immediately load the exosome suspension onto a chromatography column filled with Sepharose 2B (*see* **Note 1**).
11. Elute exosomes and proteins with PBS. Collect 1 ml per one fraction.

12. Perform identification of exosomes in the enriched fractions (usually fractions 3–4) using Western Blot analysis (*see* **Note 9**).
13. Lyse exosomes with the use of exoLB in the ratio of 2:1 (v/v) and incubate the solution at 95 °C for 5 min. Perform protein assay with the use of tryptophan fluorescence method.

3.2 Protein/Peptide Assay by Tryptophan Fluorescence Method

A versatile method of protein/peptide assay is needed at different stages of the protocol: from extraction of exosomal proteins to mass spectrometry analysis. For this purpose we adapted the tryptophan fluorescence method as proposed by Wiśniewski and Gaugaz [13]. A set of standards of L-tryptophan is prepared based on the assumption that the content of tryptophan in animal tissues is 1.17% and taking into account the buffer proteins/peptides are dissolved in, i.e., the sample matrix (*see* **Notes 10** and **11**).

1. Prepare a set of standards of L-tryptophan.
2. Load both standard solutions and protein/peptide samples into the wells of a selected micro-well plate (for sample volume below 50 μl you can use a 384-well plate) and measure tryptophan fluorescence in the conditions listed below:

   Excitation: 295 nm, 5 nm bandwidth.

   Emission: 350 nm, 20 nm bandwidth.

   Temperature: 25 °C.

   Top optic.

   Individual measurements: 30 reads, 50 μs integration time.

   Before each measurement: orbital-type shaking for 5 s followed by 2 s resting time.

   Z-position to be set manually (e.g., 18,000 μm).

   Detector gain to be set manually (e.g., for the most concentrated standard).
3. Perform at least three measurement series. Construct a calibration graph plotting fluorescence vs protein/peptide concentration or total protein/peptide content. The calibration dependence should be linear. Determine protein/peptide concentration in your sample from the calibration equation, always taking into account the limit of quantification (LOQ) of the method calculated as:

$$\mathrm{LOQ} = \frac{10 \times \mathrm{SD_{bl}}}{S}$$

   where $\mathrm{SD_{bl}}$ is the standard deviation for a blank and S is the slope of a calibration plot.
4. Transfer each sample from a micro-well to a test tube for further processing.

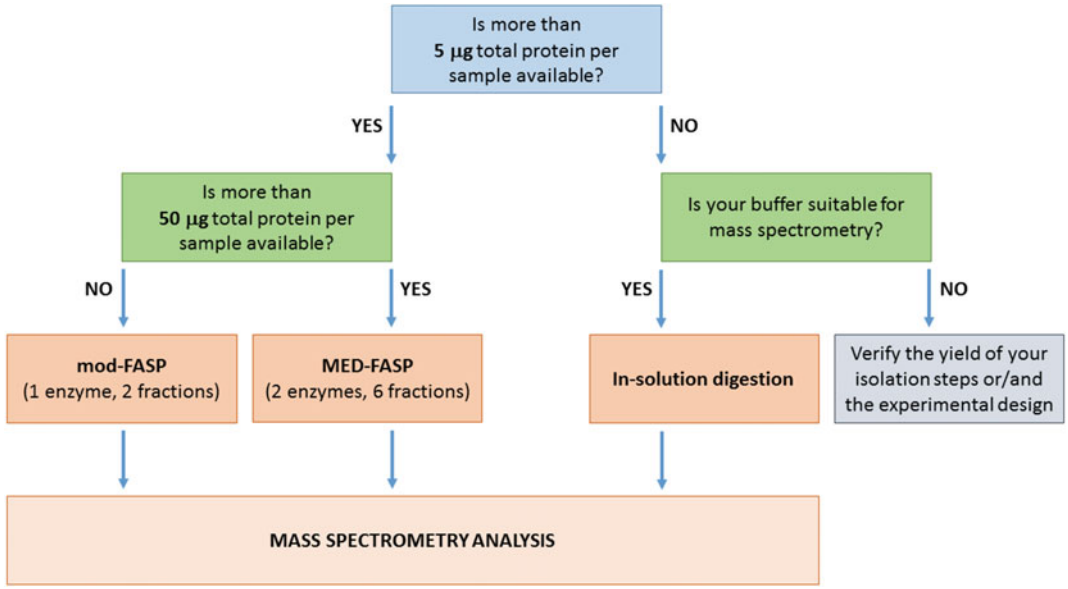


Fig. 2 Possible workflows of sample preparation for mass spectrometry analysis of exosomal proteins

3.3 Protein Digestion and Fractionation

Depending on the result of a protein assay and the kind of a buffer/solution the proteins are dissolved in, three procedures can be performed (Fig. 2). In our protocol protein extracts are subjected to a modified version (mod-FASP) of the multiple-enzyme digestion filter-aided sample preparation procedure (MED-FASP) proposed by Wiśniewski et al. [9, 10]. The latter procedure will not be described here and interested readers should refer to the original papers and educational materials provided by these authors [11, 12].

3.3.1 Modified Filter-Aided Sample Preparation (Mod-FASP)

1. Prepare tip columns: you will need one SAX-tip column and two C18-tip columns per sample (*see* **Note 12**).
2. Load up to 50 µl of exosomal protein extract and 200 µl of 8 M urea in 0.1 M Tris–HCl pH 8.5 into a Microcon spin ultrafiltration unit. Centrifuge for 15 min at 14,000 × *g* (room temperature, RT).
3. Add 200 µl of 8 M urea in 0.1 M Tris–HCl pH 8.5 (100 µl), centrifuge for 15 min at 14,000 × *g* (RT).
4. Add 50 µl of iodoacetamide solution (50 mM), mix in a shaker for 1 min at 600 rpm (RT). Incubate in darkness for 20 min. Centrifuge for 15 min at 14,000 × *g* (RT).
5. Add 100 µl of 8 M urea in 0.1 M Tris–HCl pH 8.5, centrifuge for 15 min at 14,000 × *g* (RT). Repeat this step twice.
6. Add 100 µl of 50 mM Tris–HCl pH 8.5, centrifuge for 15 min at 14,000 × *g* (RT). Repeat this step twice.

7. Replace the collection tube with a new one. Add 40 µl of 50 mM Tris–HCl pH 8.5 with trypsin (enzyme-to-protein ratio of 1:100, w/w), mix in a shaker for 1 min at 600 rpm. Incubate in a humid chamber at 37 °C for 18 h (*see* **Note 5**).
8. Centrifuge the filter units for 15 min at 14,000 × *g* (RT).
9. Add 160 µl water and centrifuge again (15 min, 14,000 × *g*, RT).
10. Dilute thus obtained tryptic peptides with 200 µl of diluted BRUB pH 5.
11. Precondition SAX-tip columns via consecutive washes with: 100 µl methanol, 100 µl 1 M NaOH, 100 µl of diluted BRUB pH 5, and again 100 µl of diluted BRUB pH 5; follow addition of each solution by centrifugation: 4000 × *g*, 5 min (RT).
12. Precondition C18-tip columns via consecutive washes with: 50 µl methanol, 50 µl 60% ACN/0.1% TFA, and 50 µl 0.1% TFA/H_2O; follow addition of each solution by centrifugation: 4000 × *g*, 5 min (RT).
13. Insert a SAX-tip column in a C18-tip column; load tryptic peptides: 2 × 200 µl, each loading followed by centrifugation: 5000 × *g*, 3 min (RT).
14. Add 100 µl of diluted BRUB pH 5, centrifuge: 5000 × *g*, 3 min.
15. Transfer the SAX-tip column to the next C18-tip column; add 100 µl of diluted BRUB pH 2, centrifuge: 5000 × *g*, 3 min (RT).
16. Discard SAX-tip column; wash C18-tip columns with 50 µl of 0.1% TFA/H_2O (centrifugation at 5000 × *g*, 3 min, RT).
17. Elute peptides from C18-tip columns with 50 µl of 60% ACN/0.1% TFA (centrifugation at 5000 × *g*, 3 min, RT).
18. Remove the elution buffer from peptide fractions in a vacuum concentrator (*see* **Note 11**). Reconstitute with water (50 µl or less if you expect low peptide content).
19. Determine peptide content in the fractions using the tryptophan fluorescence method (refer to Subheading 3.2).
20. Before LC-MS/MS analysis acidify peptide fractions with 1% TFA/H_2O (v/v) to reach the final TFA concentration of ca. 0.1% (v/v).
21. Dilute pH 5 peptide fraction with 0.1% TFA in order to achieve equal peptide concentration in both fractions. Equal fraction volumes need to be loaded onto an LC-column for each sample to maintain constant measurement conditions.

3.3.2 In-Solution Digestion

This protocol is adapted to a total sample volume of up to 10 μl. The final reaction mixture volume is 30 μl.

1. Mix 15 μl of 50 mM NH_4HCO_3 and 1.5 μl of 100 mM DTT in a 0.5 ml centrifuge tube.
2. Add the protein solution and adjust the final volume of the mixture to 27 μl with ultrapure water.
3. Incubate the mixture at 95 °C for 5 min, cool down to room temperature.
4. Add 3 μl of 100 mM IAA to the tube and incubate in the dark at room temperature for 20 min.
5. Add a proper volume of 0.1 μg/μl trypsin to the mixture to reach enzyme:protein ratio of 1:100 w/w (e.g., 1 μl for 10 μg of protein), incubate at 37 °C for 18 h.
6. Terminate the reaction by adding 1.5 μl 10% TFA.

3.4 Mass Spectrometry

Instrument settings given below should be considered only as a starting point for your MS method development. The optimal measurement conditions for your system may vary and should be adjusted according to your samples (*see* **Notes 13–17**). Depending on the peptide concentration in protein digests one should employ a sufficiently sensitive mass spectrometer (e.g., quadrupole-Orbitrap for the total peptide content below 5 μg). Nevertheless, if contaminants (e.g., high abundant proteins) are expected in a sample, application of a highly sensitive system may have the opposite effect to the intended one. Strictly speaking, increase in sensitivity of an MS system may not result in increase in the number of identified proteins.

3.4.1 LC-MALDI MS/MS

1. Nano-LC conditions:

   Buffer A: 0.05% TFA/H_2O.

   Buffer B: 90% ACN, 0.05% TFA.

   Pre-column: C18, 100 μm × 2 cm, 5 μm granulation, 100 Å.

   Analytical column: C18, 75 μm × 15 cm, 3 μm granulation, 100 Å.

   Acetonitrile gradient: from 2 to 45%, in 0.05% TFA.

   Flow rate: 300 nl/min (113 min).

2. Fraction collection on a MALDI target plate:

   MTP AnchorChip 1536 target plate.

   α-Cyano-4-hydroxycinnamic acid (HCCA) solution (*see* **Note 18**).

   Eluate from the analytical column is mixed with HCCA solution and spotted over 680 fractions on an MTP AnchorChip 1536 target plate.

3. MS and MS/MS conditions:
   (a) Collect MS spectra in positive reflector mode within tryptic peptide range (800–4000 *m/z*), 3000 shots from each LC fraction, random walk activated.
   (b) Fragment ions with S/N higher than 10, sum up 5000 shots for a fragment (MS/MS) spectrum.
4. Database search:
   (a) Use a selected program for database search and protein identification—we recommend Mascot Server 2.5.1 (Matrix Science, London, UK) and ProteinScape 3.1 (Bruker).
   (b) Set up proper search conditions: e.g., Swiss-Prot human database with a precision tolerance of 50 ppm for peptide masses and 0.5 Da for fragment ion masses; one missed cleavage; select Carbamidomethyl (C) and Oxidation (M) as fixed and variable modifications, respectively.
   (c) When using ProteinScape software you can perform protein list compilation by ProteinExtractor: ions score cutoff, peptide rank cutoff, and minimum peptide length set at: 15.0, 10, and 5, respectively; identity score calculated by the search engine.

3.4.2 Quadrupole-Orbitrap LC-MS/MS

1. Nano-LC conditions:

   Buffer A: 0.1% formic acid/H_2O.

   Buffer B: 90% ACN, 0.1% formic acid.

   Loading buffer for trapping: 98% H_2O, 0.1% TFA.

   Trapping column: C18, 300 μm × 5 mm, 5 μm granulation, 100 Å.

   Analytical column: C18, 75 μm × 25 cm, 2 μm granulation, 100 Å; 30 °C.

   Acetonitrile gradient: from 4% to 60%, in 0.1% formic acid.

   Flow rate: 300 nl/min (230 min).

2. MS and MS/MS conditions:

   Data-dependent MS/MS mode with survey scans acquired at resolution of 70,000 at *m/z* 200 in MS mode, and 17,500 at *m/z* 200 in MS2 mode.

   Scanning *m/z* range of 300–2000, positive ion mode.

   Higher energy collisional dissociation (HCD) ion fragmentation with normalized collision energies set to 25.

3. Database search:

   Perform protein identification using a selected database, e.g., Swiss-Prot human database with a precision tolerance

10 ppm for peptide masses and 0.08 Da for fragment ion masses. Use a selected software for estimation of abundances of identified proteins, e.g., MaxQuant 1.4.1.1 software or Proteome Discover 2.0 for Thermo raw files.

4 Notes

1. While maintaining proper dimensions of the chromatographic bed (i.e., height and diameter), a wide range of products can serve as a column for Sepharose 2B. A mini-column suitable for the described conditions can be purchased (i.e., Econo Pac chromatography columns, Bio-Rad) [14]. In order to avoid bed leakage, a frit made of chemically inert material is placed at the outlet of the column. A mini-column is filled with Sepharose 2B, a second frit is placed on the top of Sepharose, and washed 2–3 times with elution buffer (i.e., PBS) until the proper bed height of 10 cm (inner diameter of 1.5 cm) is reached. Bed drying and bed leakage should be avoided.
2. We prepare tryptophan stock solution every week and store it at 4 °C; however, freshly prepared solution can also be aliquoted and stored at −20 °C for 3 months. Working solutions (0.1 mgTrp/ml and 0.01 mgTrp/ml in water) should be prepared every day before use.
3. Add 40 μl of the Trypsin Resuspension buffer provided along with the enzyme (or 50 mM acetic acid) into the glass vial with enzyme lyophilisate (20 μg), vortex mixture thoroughly, and spin down. Divide the obtained trypsin stock solution into aliquots of 10 μl and store at −20 °C for up to 6 months. Avoid multiple thaw-freeze cycles. Enzymatic activity of trypsin is reversibly blocked only in acidic conditions; therefore, trypsin solutions in water or any digestion buffer (e.g., 25 mM ammonium bicarbonate) should always be used freshly.
4. The quality of plastic consumables is of prime importance when samples are to be analyzed by mass spectrometry. Always use materials of the highest quality. We recommend Safe-Lock Tubes and epT.I.P.S.® by Eppendorf AG (Hamburg, Germany). Avoid all kinds of low-bind or siliconized plastics. For long storage of organic solvents borosilicate glass vessels or PTFE-coated plastic bottles are highly recommended instead of polypropylene ones.
5. A suitable plastic laboratory box (e.g., 25 cm × 17 cm × 10 cm, L × W × H), equipped with a well-fitting lid, can serve as a humid chamber. Put several sheets of paper towel or cellulose wadding on the bottom of the box and wet them well with water, pour off the excess of water. Place a small tube rack

inside. Make sure that the rack does not cover the entire surface of the wetted cellulose. Pre-heat the humid chamber in a laboratory incubator (37 °C). Put a collection tube with a Microcon centrifuge filter in the rack and open the lid of the collection tube. Close the box and leave it in the incubator for 18 h.

6. Take as many cells as necessary to achieve proper confluency of cells at the moment of medium collection. We use two 150 cm^2 cell culture flasks per an experimental point (the final volume of the conditioned medium is 50 ml). The number of cells is closely related to a cell line, usually we seed between 2×10^6–5×10^6 adherent cells per 150 cm^2 in 25 ml of medium. Harvest your cells when they reach 70–80% confluency for adherent cells, or 60–70% of their maximum concentration for cells grown in a suspension.
7. Depending on the type of the rotor we suggest: (1) in the case of a swing bucket it is better to collect the supernatant with the use of a Pasteur pipette, since the pellet is localized right on the bottom of the centrifugation tube, it is poorly adherent; (2) in the case of a fixed angle rotor it is better to decant the supernatant in a single motion, since the pellet is localized on a side wall of a tube. This remark is important in subsequent steps of the procedure—in fact it determines the proceeding during the isolation process.
8. Directly before use, concentrators are washed with 50 ml of phosphate-buffered saline solution to remove possible postproduction impurities. The solution is then discarded. In order to purify the membrane from glycerine which it is originally covered with, the concentrator is filled with PBS again, centrifuged for 5 min, 700 $\times g$ and 4 °C. One should avoid drying out of the membrane. Separation should be performed directly after conditioning of the concentrator.
9. Assessment of quantity and quality of exosomes in samples can be performed using five independent methods: (1) Transmission electron microscopy (TEM) techniques: Coat exosomes with 0.125% (w/v) Formvar in chloroform on copper grids. Grids can be stained with 1% (w/v) uranyl acetate in doubly distilled H_2O. (2) Protein quantification using a BCA protein assay kit or tryptophan fluorescence measurement. (3) Lipid quantification using a proper reagents kit. (4) Tunable Resistive Pulse Sensing (TRPS) (i.e., qNano by Izon) for size distribution and concentration of particles (according to the manufacturer's instructions). (5) Western Blot: Perform SDS gel electrophoresis with equal amounts (min. 5–20 μg protein) of each exosomal fraction followed by Western blotting and detection of antigens of interest.

Table 1
Composition of standard solutions of L-tryptophan

Standard Solution No.	Buffer[a] (μl)	Trp solution 0.01 μg/μl (μl)	Water (μl)	Total Trp content (μg)	Total protein/peptide content (μg)	Protein/peptide concentration (μg/μl)
1	80	0	20	0.00	0.0	0.000
2	80	4	16	0.04	3.4	0.034
3	80	12	8	0.12	10.3	0.103
4	80	20	0	0.20	17.1	0.171

Standard Solution No.	Buffer[a] (μl)	Trp solution 0.10 μg/μl (μl)	Water (μl)	Total Trp content (μg)	Total protein/peptide content (μg)	Protein/peptide concentration (μg/μl)
5	80	4	16	0.40	34.0	0.34
6	80	12	8	1.20	103.0	1.03
7	80	20	0	2.00	171.0	1.71

[a]exoLB in the case of protein extracts; a mixture of 50 mM Tris–HCl pH 8.5 and water, 1:4 (v/v) in the case of tryptic digests; water in the case of peptide fractions

10. An exemplary way of preparation of tryptophan standard solutions is presented in Table 1. Although the tryptophan fluorescence method is compatible with many popular solutes employed in protein lysis/extraction buffers (for details refer to [13]), tryptophan standards must reproduce the real sample matrix (i.e., the buffer the assayed proteins/peptides are dissolved in) as accurately as possible.
11. Before tryptophan fluorescence measurement the elution buffer in peptide fractions (60% ACN, 0.1% TFA) should be replaced with water, since the presence of acetonitrile and trifluoroacetic acid in such concentrations results in decrease of the method sensitivity by the factor of 4 in comparison to water.
12. Cut the lid of a 2 ml reaction tube (two intersecting incisions) and insert a 200 μl pipette tip in the obtained slit, press firmly. Cut six plugs of a strong anion exchanger (SAX) extraction filter with the use of a blunt needle, transfer the needle to the pipette tip and push the plugs out of the needle with the use of a plastic or metal wire. Press the material tightly in the tip—the plugs should not separate one from another when in use, nevertheless too tight packing may block a tip column. Repeat

the procedure with C18 extraction filter, but use three plugs instead of 6.

13. Avoid contamination during sample preparation. Always use gloves and proper lab coats. It is easy to contaminate a sample, e.g., with keratins or other abundant proteins when touching lab equipment or sample tubes without gloves. When using gloves be careful not to touch things like door handles, computer mouse or keyboard etc. If still keratins are overrepresented in an identification report (MASCOT), check for other contamination sources like dust in the lab and prepare your samples under a hood. Other contamination sources are polyethylene glycols (PEGs) and plasticizers like phthalate derivatives. Use only high quality plastics or glass (*see* **Note 4**). PEGs are often introduced when using detergents in your lab.
14. For peptides separation using LC/MS, different column configurations may be alternatively adapted. The best configuration is as described here, using a pre-column (trap column) prior to the proper separation step, but also direct column separation may be applied. It should be noted here that total capacity of a column and a pre-column should be investigated to avoid their overloading.
15. This procedure can also be adapted to peptide labeling approaches like iTRAQ technique instead of the described label-free approach. If applicable, several remarks should be considered:
    (a) Not all MS systems are compatible with peptide isobaric labeling on MS2, for example, ion trap mass spectrometers are limited in registration of all fragment ions, reporter ions may be missed in this case.
    (b) When using QExactive instrument remember to set up "first mass" measurement at 100 m/z for acquisition of all MS2 spectra.
    (c) When using labeling at MS stage (i.e., SILAC, ICAT), remember that sample complexity is doubled which may cause sensitivity or suppression problems.
16. This procedure may also be adapted to protein modifications analysis (PTMs). Just remember to include proper modifications and their sites as variable modifications in MASCOT search parameters. Avoid searching too many modifications at a time as it can give rise to unspecific peptide identification. When dealing with modifications occurring at lysine (or arginine) residues consider increasing the number of possible miscleavages.
17. In a label-free approach, samples from a given experiment should be analyzed in one batch when possible (in a

randomized order). Proper calibration of a mass spectrometer should be performed before the first analysis and after every 10 or 15 samples. If applicable, use quality control samples (QC) between selected runs. All these remarks are important for proper further data formatting and preparation prior to statistical analysis (data normalization and alignment).

18. Load internal syringe of the fraction collector with a matrix solution consisting of:
    (a) For nano-LC fractions:
        - 748 μl of 95% ACN in water, 0.1% TFA.
        - 36 μl of HCCA saturated in 90% ACN in water, 0.1% TFA.
        - 8 μl of 10% trifluoroacetic acid in water.
        - 8 μl 100 mM $NH_4H_2PO_4$ in water.
    (b) For external calibrants:
        - 748 μl of 85% ACN in water, 0.1% TFA.
        - 36 μl of HCCA saturated in 90% ACN in water, 0.1% TFA.
        - 8 μl 10% trifluoroacetic acid in water.
        - 8 μl 100 mM $NH_4H_2PO_4$ in water.

Acknowledgements

This work was supported by the National Science Centre, Poland, Grant 2013/11/B/NZ7/01512.

References

1. Théry C, Clayton A, Amigorena S et al (2006) Isolation and characterization of exosomes from cell culture supernatants and biological fluids. Curr Protoc Cell Biol 30:3.22.1–3.22.29
2. Tickner JA, Urquhart AJ, Stephenson SA et al (2014) Functions and therapeutic roles of exosomes in cancer. Front Oncol 4:127
3. Zitvogel L, Regnault A, Lozier A et al (1998) Eradication of established murine tumors using a novel cell-free vaccine: dendritic cell-derived exosomes. Nat Med 4:594–600
4. Andre F, Schartz NE, Movassagh M et al (2002) Malignant effusions and immunogenic tumour-derived exosomes. Lancet 360:295–305
5. Ginestra A, Miceli DL, Dolo V et al (1999) Membrane vesicles in ovarian cancer fluids: a new potential marker. Anticancer Res 19:3439–3445
6. Adams M, Navabi H, Croston D et al (2005) The rationale for combined chemo/immunotherapy using a Toll-like receptor 3 (TLR3) agonist and tumour-derived exosomes in advanced ovarian cancer. Vaccine 23:2374–2378
7. Nilsson J, Skog J, Nordstrand A et al (2009) Prostate cancer-derived urine exosomes: a novel approach to biomarkers for prostate cancer. Br J Cancer 100:1603–1607
8. Abramowicz A, Widlak P, Pietrowska M (2016) Proteomic analysis of exosomal cargo: the challenge of high purity vesicle isolation. Mol Biosyst 12:1407–1419
9. Wiśniewski JR, Zougman A, Nagaraj N et al (2009) Universal sample preparation

method for proteome analysis. Nat Methods 6:359–363

10. Wiśniewski JR, Zougman A, Mann M (2009) Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. J Proteome Res 8:5674–5678
11. Filter Aided Sample preparation (FASP) Method. http://www.biochem.mpg.de/226356/FASP. Accessed 7 Jun 2016
12. Tutorials. http://www.biochem.mpg.de/226455/Tutorials. Accessed 7 Jun 2016
13. Wiśniewski JR, Gaugaz FZ (2015) Fast and sensitive total protein and peptide assays for proteomic analysis. Anal Chem 87:4110–4116
14. Hong CS, Funk S, Muller L et al (2016) Isolation of biologically active and morphologically intact exosomes from plasma of patients with cancer. J Extracell Vesicles 5:29289

Part V

From Genotype to Phenotype

Chapter 23

Virus-Induced Gene Silencing (VIGS) and Foreign Gene Expression in *Pisum sativum* L. Using the "One-Step" *Bean pod mottle virus* (BPMV) Viral Vector

Chouaïb Meziadi, Sophie Blanchet, Valérie Geffroy, and Stéphanie Pflieger

Abstract

Plant viral vectors have been developed to facilitate gene function studies especially in plant species not amenable to traditional mutational or transgenic modifications. In the *Fabaceae* plant family, the most widely used viral vector is derived from *Bean pod mottle virus* (BPMV). Originally developed for over-expression of foreign proteins and VIGS studies in soybean, we adapted the BPMV-derived vector for use in other legume species such as *Phaseolus vulgaris* and *Pisum sativum*. Here, we describe a protocol for efficient protein expression and virus-induced gene silencing (VIGS) in *Pisum sativum* leaves and roots using the "one-step" *Bean pod mottle virus* (BPMV) viral vector.

Key words Garden pea, *Bean pod mottle virus*, Virus-induced gene silencing (VIGS), Functional genomics, RNAi, Post-transcriptional gene silencing, Legume

1 Introduction

Plant viruses, mainly positive-sense RNA plant viruses, have been engineered as recombinant viral vectors either to express foreign proteins in plant cells or to down-regulate expression of targeted plant genes [1, 2]. Indeed, viruses are obligate pathogens that make their whole life cycle inside their host cells. After decapsidation of their particles in an infected cell cytoplasm, the RNA viral genome is recognized by the cell translation machinery and is translated in one or several viral proteins [3]. Thus, insertion of a foreign open reading frame (ORF) at an adequate position within the viral genome leads to transient expression of this protein. Therefore, virus-based vectors began to be used to produce commercial products like high-value pharmaceutical proteins, crude preparation of enzymes for industrial use, as well as vaccine antigens and antibodies for medical applications [1, 4]. On the other hand, secondary

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_23,

structures of the single-stranded RNA genome molecule as well as double-stranded RNA molecules generated during viral replication (*i.e.*, replicative forms) induce a natural RNA-silencing defense response related to post-transcriptional gene silencing (PTGS) [5]. Plant infection by a recombinant virus carrying a fragment of a targeted endogenous gene will induce the RNA silencing pathway and sequence-specific degradation of mRNA corresponding to the target gene. Thus expression of the target gene will be down-regulated [5]. This technology was called virus-induced gene silencing (VIGS). VIGS is usually used in plant species that are not readily amenable to stable genetic transformation like the *Fabaceae* plant family, for example [6, 7]. VIGS is rapid, feasible in different genetic backgrounds provided that the genotype is susceptible to the selected viral strain.

In the *Fabaceae* plant family, the most widely used viral vector is derived from *Bean pod mottle virus* (BPMV, genus *Comovirus*). BPMV is a positive-strand RNA virus that was first discovered in common bean, but was subsequently shown to infect many other legume species such as soybean and pea [8, 9].

The genome of BPMV is bipartite, with two RNA molecules RNA1 (~6 kb) and RNA2 (~3.6 kb) that are encapsidated in separate isometric particles. RNA1 and RNA2 are expressed as polyproteins that are subsequently processed by proteinases for the synthesis of mature viral proteins. Three generations of BPMV-derived vectors have been successively developed with the aim of increasing the potential of BPMV as a viral vector for functional genomics (reviewed in [10]). In all three vectors, insertion of foreign DNA fragments for gene expression and/or VIGS induction is made in RNA2. The third-generation BPMV-derived vector, designed in soybean by Zhang et al. [11, 12], presents important improvements compared to previous generations. First, cloning of foreign sequences into BPMV RNA2 is facilitated by the introduction of a *Bam*H1 restriction site after the translation stop codon of RNA2. Second, delivery of the BPMV vector into plants is possible *via* direct DNA rubbing of infectious plasmid DNA, a procedure adapted to high-throughput studies. Third, this BPMV vector is derived from the IA-Di1 isolate which induces very mild visual symptoms on infected soybean plants, thus avoiding possible interference between viral symptoms and silencing phenotypes [11]. All these improvements make this new BPMV vector an ideal "one-step" viral vector (so-called because there is no need for in vitro transcription, *Agrobacterium* transformation, or coating to gold particles for biolistic delivery). This vector is adapted to high-throughput genomic studies and has enabled efficient, cost-effective, and simplified functional screening of genes in soybean [11].

Recently, we adapted the "one-step" BPMV vector for gene expression and VIGS induction in *Phaseolus vulgaris* and *Pisum*

sativum, two agronomical important legume species for which complete genome sequences are available [13–15]. In our work, we optimized a protocol for rub-inoculation of infectious BPMV-derived plasmids in *P. vulgaris* cv. Black Valentine, a highly susceptible genotype to BPMV [11, 16]. This delivery procedure is rapid, cheap, and accessible to every laboratory (*see* [17] for illustrated protocol). Primary infection rates range from 55% to 91% *in P. vulgaris* [16]. We optimized a protocol of mechanical inoculation using infected leaf sap, derived from *P. vulgaris*-infected plants, to inoculate healthy pea plants (*i.e.*, secondary inoculation) [18]. We obtained efficient gene expression (using the *GFP* gene as a foreign gene) and VIGS induction (using the *PDS—phytoene desaturase*—and *Korrigan-1* genes as reporter genes) in both leaves and roots.

This paper describes in detail the protocol to perform gene expression and VIGS assays in *P. sativum* using the "one-step" BPMV-derived vector.

2 Materials

2.1 Plant Material and Growth Conditions

1. *Pisum sativum* (*P. sativum*) seeds. Seeds of pea genotypes were obtained from INRA Dijon (France). As a control genotype, we recommend the use of *P. sativum* cv. "Champagne," the most susceptible genotype tested with BPMV (*see* **Note 1**).
2. *Phaseolus vulgaris* (*P. vulgaris*) cv. Black Valentine seeds. Seeds can be obtained from the Centro Internacional de Agricultura Tropical (CIAT, Colombia). As cv. "Black Valentine" is highly susceptible to BPMV, we recommend this genotype for primary inoculation to generate infected leaf sap used for further inoculation of any other pea genotype.
3. Vermiculite 1–4 mm (Agrena, Rungis, France).
4. Plastic pots (7 × 7 × 6 cm pots and 20 cm diameter pots of 4 L).
5. Greenhouse or growth chamber for plant growth at 23 °C, 16 h light/8 h dark cycle, 70% relative humidity.
6. Dark room at 20 °C (for plant stocking 24 h prior inoculation).
7. Greenhouse or growth chamber for phytopathological tests at 19 °C, 16 h light/8 h dark cycle, 70% relative humidity (*see* **Note 2**).
8. Hydroponic culture system for making VIGS in roots.

2.2 Primary Inoculations of Phaseolus vulgaris to Produce Viral Inoculum

1. Recombinant BPMV RNA1 plasmid (pBPMV-IA-R1M) [11] (*see* **Note 3**).
2. Recombinant BPMV RNA2 plasmid (pBPMV-IA-V1, as a BPMV wild-type control) [11] (*see* **Note 3**).

3. Recombinant BPMV RNA2 plasmid (pBPMV-GFP2) (green fluorescent protein) (as a gene expression positive control) [11] (*see* **Note 3**).
4. Recombinant BPMV RNA2 plasmid, (pBPMV-IA-PsPDS336bp) (phytoene desaturase) (as a VIGS positive control in leaves) [18] (*see* **Note 3**).
5. Recombinant BPMV RNA2 plasmid, (pBPMV-IA-PsKOR-345bp or pBPMV-IA-PsKOR-470bp) (endo-1,4-β-Glucanase) (as a VIGS positive control in roots) [18] (*see* **Note 3**).
6. Plasmid Maxi Kit (QIAGEN, Hilden, Germany).
7. Buffer: 0.05 M potassium phosphate pH 7.0.
8. Carborundum 0.037 mm (used as an abrasive).
9. Absorbant paper.

2.3 Secondary Inoculations of Pisum sativum *Using Infected Leaf Tissues from* Phaseolus vulgaris

1. Fresh, dried, or frozen leaf tissues infected by BPMV-0 (pBPMV-IA-R1M + pBPMV-IA-V1).
2. Fresh, dried, or frozen leaf tissues infected by BPMV-GFP (= pBPMV-IA-R1M + pBPMV-GFP2) as a gene expression positive control.
3. Fresh, dried, or frozen leaf tissues infected by BPMV-PsPDS (= pBPMV-IA-R1M + pBPMV-IA-*Ps*PDS336bp) as a VIGS positive control in leaves.
4. Fresh, dried, or frozen leaf tissues infected by BPMV-PsKOR1 (= pBPMV-IA-R1M + pBPMV-IA-PsKOR-345bp or pBPMV-IA-PsKOR-470bp) as a VIGS positive control in roots.
5. Buffer: 0.05 M potassium phosphate pH 7.0.
6. Mortar (9 cm diameter) and pestle (12 mm diameter).
7. Scalpel and sterile blades.
8. Carborundum 0.037 mm (used as an abrasive).
9. Miracloth.
10. Absorbant paper.

3 Methods

3.1 Primary Inoculations of Phaseolus vulgaris *to Produce BPMV Viral Inoculum*

1. To produce BPMV inoculum for *P. sativum* inoculations, mechanical inoculation of BPMV vectors by direct DNA rubbing of BPMV-derived infectious plasmids was achieved in *P. vulgaris* cv. "Black Valentine," a highly susceptible genotype. The procedure was described previously [17].

3.2 Secondary Inoculations of Pisum sativum Using Infected Leaf Tissues from Phaseolus vulgaris

3.2.1 VIGS and Foreign Gene Expression in Pisum sativum Leaves

1. Sow seeds of *P. sativum* in vermiculite in plastic pots (7 × 7 × 6 cm pots) and grow in a growth chamber at 23 °C under a 16 h light/8 h dark cycle and 70% relative humidity.
2. Seedlings with two fully expanded leaves are ready for inoculation (after 10–12 days in our conditions) (*see* Fig. 1a). At this stage, transplant three seedlings in moist vermiculite in a 20 cm plastic pot.
3. Place the plants in a dark room for 24 h prior to inoculation.
4. To prepare the inoculum from the infected leaf tissues of *P. vulgaris*, put a fresh or a frozen infected leaflet of common bean "Black Valentine" in a mortar.
5. Grind briefly the tissue with a pestle to obtain a green mash.
6. Add ~3–4 mL of 50 mM potassium phosphate buffer, pH 7 to make leaf sap.
7. Grind again with the pestle to obtain a green leaf sap. As leaf sap is usually heterogeneous, let it settle a few minutes.
8. Use a sterile scalpel blade to scarify the upper leaf and stipules surfaces (*see* Fig. 1b). This operation is essential to make superficial incisions in the waxy layer and in the upper epidermis of the leaf so that viral infection is optimal. Depending on the size of the leaflet surface, four to six parallel incisions are made on one leaflet (*see* Fig. 1b).
9. Powder the wounded upper surface of the two first leaves and their stipules with carborundum. Don't dust too much

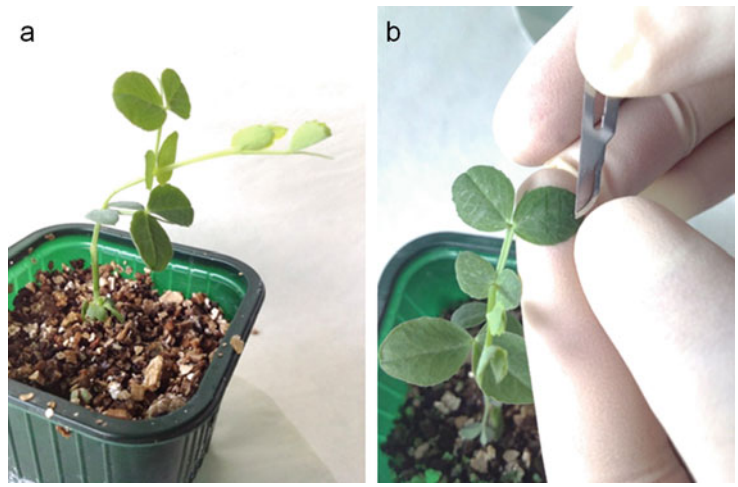


Fig. 1 Secondary inoculations of *Pisum sativum* using the "one-step" *Bean pod mottle virus* (BPMV) vector. (**a**) For optimal inoculations, use seedlings at the two fully expanded leaf stage (10–12 days-old seedlings in our conditions). (**b**) Scarification step of *Pisum sativum* leaves and stipules: use a sterile scalpel blade to scarify the upper leaf and stipules surfaces. Depending on the size of the leaflet surface, four to six parallel incisions are made on one leaflet

carborundum because it can generate undesirable necrotic areas after rubbing.

10. Cut a Miracloth piece of ~8 × 6 cm and fold it in four.
11. Soak the folded Miracloth in the leaf sap.
12. Rub gently the leaf surface with the folded Miracloth. All the leaf surface should be rubbed.
13. At ~3–4 min after rubbing, rinse the inoculated leaflets and stipules with tap water contained in a wash bottle until all carborundum is removed.
14. Remove the excess of water using absorbant paper.
15. Place the inoculated plants in a growth chamber at 19 °C under a 16 h light/8 h dark cycle and 70% relative humidity.
16. Fertilize plants after mechanical inoculation by pouring the nutritive solution directly in the pot saucer (approximately 500 mL in a 20 cm diameter saucer). Fertilize at a 3–4 days interval (*see* **Note 4**).
17. Viral symptoms induced by BPMV-0 occur on the upper systemic leaves at about 2–4 weeks postinoculation depending on the genotype tested (in [18] *see* Fig. 2).
18. Green fluorescence by BPMV-GFP (= pBPMV-IA-R1M + pBPMV-GFP2) can be seen on the inoculated leaves at about 7–14 days post-inoculation under UV light and after 3–4 weeks post-inoculation on the upper systemic leaves (in [18], *see* Fig. 1 and S1).
19. White photobleaching phenotype corresponding to *PsPDS* gene silencing induced by BPMV-PsPDS (=pBPMV-IA-R1M + pBPMV-IA-PvPDS336bp) is generally observed at 6 weeks post-inoculation in cv. "Champagne" (in [18] *see* Fig. 2), generally on the young upper leaves.
20. For all BPMV vectors, the successful infection rate of secondary inoculation is 100% in cv. "Champagne" [18]. The infection rates can decrease to 30% in other genotypes [18].

3.2.2 VIGS and Foreign Gene Expression in Pisum sativum Roots

1. Sow seeds of *P. sativum* in vermiculite in plastic pots (7 × 7 × 6 cm pots) and grow in a growth chamber at 23 °C under a 16 h light/8 h dark cycle and 70% relative humidity.
2. Seedlings with two fully expanded leaves are ready for inoculation (after 10–12 days in our conditions) and are placed in a dark room 24 h prior to inoculation.
3. Before inoculation, place the pea seedlings for *PsKOR1* silencing assays individually in hydroponic culture and use a nutrient solution as substrate (*see* **Note 4**).
4. Inoculate plants as described in the previous section.

5. After inoculation with recombinant BPMV vectors, place the hydroponic-cultured plants in a growth room at 19 °C under a 16 h light/8 h dark cycle under a humidity of 70%. Refill the level of nutrient solution regularly.
6. At 14 dpi, cut all roots of each hydroponic-cultured plant to approximately 3 cm with a sterile scalpel blade. This allows root growth to reinitiate.
7. Green fluorescence produced by BPMV-GFP (=pBPMV-IA-R1M + pBPMV-GFP2) can be seen in roots at about 5–6 weeks postinoculation using an epifluorescent microscope (in [18] *see* Fig. 3).
8. Phenotypes of root dwarfing corresponding to *PsKOR1* gene silencing induced by BPMV-PsKOR1 (= pBPMV-IA-R1M + pBPMV-IA-PsKOR-345bp or pBPMV-IA-PsKOR-470bp) is generally observed at 4 weeks post-inoculation in cv. "Champagne" (in [18] *see* Fig. 4).
9. For all BPMV vectors, the successful infection rate in cv. "Champagne" is close to 100% [18].

4 Notes

1. We recommend the use of the pea cv. "Champagne" for VIGS and gene expression experiments because this genotype is highly susceptible to BPMV in our growth conditions. For secondary inoculations using infected leaf tissues, the choice of pea genotypes may depend on your research goals. Keep in mind that the use of BPMV vectors for gene expression or VIGS requires susceptibility of the selected pea genotype to BPMV. In our work, we tested 43 cultivated pea genotypes for their susceptibility to BPMV using the BPMV-GFP construct [18]. We showed that in some susceptible pea genotypes (*e.g.*, cv. "Enduro," "James"), the GFP fluorescence is only visible in the inoculated primary leaf but the viral vector does not move to the upper non-inoculated leaves and thus no GFP fluorescence is observed in these leaves [18].
2. Keep in mind that BPMV is a viral pathogen and that its manipulation must be in accordance with your country legislation in regard to biosafety concern and containment to preclude uncontrolled virus transmission.
3. All recombinant plasmids derived from BPMV RNA1 or BPMV RNA2 are carried by recombinant *Escherichia coli* strains and plasmids were prepared as concentrated "maxi-preps" using the Qiagen kit "Plasmid Maxi Kit" according to the supplier's instructions. To construct new BPMV vectors

containing foreign fragments of interest, refer to the detailed protocol in [12].

4. Fertilize using a nutritive solution: Fertilizer Plant-Prod 14–12-32 (final concentration = 0.28 kg/L), and Fertilizer Fertiligo L (final concentration = 4.35 mL/L), tap water.

Acknowledgements

This protocol was developed and optimized for *Pisum sativum* by modifying the procedure used for BPMV VIGS in leaves of *Phaseolus vulgaris* [16, 17] and for PEBV VIGS in roots of *Pisum sativum* [19]. We thank Chunquan Zhang and Steven Whitham at Iowa State University (USA) for sharing the set of BPMV VIGS vectors and providing the common bean cv. "Black Valentine" seeds. CM was supported by fellowships from the French Research Ministry. This work was supported by grants from Institut National de la Recherche Agronomique, Centre National de la Recherche Scientifique, Université Paris-Sud, LabEx Saclay Plant Sciences-SPS (ANR-10-LABX-0040-SPS), the 3P project (Plant Phenotyping Platform), and the PeaMUST project (ANR-11-BTBR-0002).

References

1. Lico C, Chen Q, Sant L (2008) Viral vectors for production of recombinant proteins in plants. J Cell Physiol 216:366–377
2. Burch-Smith TM, Anderson JC, Martin GB, Dinesh-Kumar SP (2004) Applications and advantages of virus-induced gene silencing for gene function studies in plants. Plant J 39 (5):734–746
3. Hull R (2014) Plant virology, 5th edn. Elsevier, Amsterdam
4. Canizares MC, Nicholson L, Lomonossoff GP (2005) Use of viral vectors for vaccine production in plants. Immunol Cell Biol 83:7–37
5. Kumagai MH, Donson J, Dellacioppa G, Harvey D, Hanley K, Grill LK (1995) Cytoplasmic inhibition of carotenoid biosynthesis with virus-derived RNA. Proc Natl Acad Sci U S A 92:1679–1683
6. Somers DA, Samac DA, Olhoft PM (2003) Recent advances in legume transformation. Plant Physiol 131:892–899
7. Svabova L, Smykal P, Griga M, Ondrej V (2005) *Agrobacterium*-mediated transformation of *Pisum sativum in vitro* and *in vivo*. Biol Plant 49:361–370
8. Giesler LJ, Ghabrial SA, Hunt TE, Hill JH (2002) *Bean pod mottle virus*: a threat to US soybean production. Plant Dis 86 (12):1280–1289
9. Brunt AA, Crabtree K, Dallwitz MJ, Gibbs AJ, Watson L, Zurcher EJ (1996) Plant viruses online: descriptions and lists from the VIDE database. Version: 20th Aug 1996. http://biologyanueduau/Groups/MES/vide/
10. Pflieger S, Richard MM, Blanchet S, Meziadi C, Geffroy V (2013) VIGS technology: an attractive tool for functional genomics studies in legumes. Funct Plant Biol 40(12):1234–1248
11. Zhang C, Bradshaw JD, Whitham SA, Hill JH (2010) The development of an efficient multipurpose *Bean pod mottle virus* viral vector set for foreign gene expression and RNA silencing. Plant Physiol 153(1):52–65
12. Zhang C, Whitham SA, Hill JH (2013) Virus-induced gene silencing in soybean and common bean. Methods Mol Biol 975:149–156
13. Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, Torres-Torres M, Geffroy V, Moghaddam SM, Gao D, Abernathy B, Barry K, Blair M, Brick MA, Chovatia M, Gepts P, Goodstein DM, Gonzales M, Hellsten U, Hyten DL, Jia G, Kelly JD, Kudrna D, Lee R, Richard MM, Miklas PN, Osorno JM,

Rodrigues J, Thareau V, Urrea CA, Wang M, Yu Y, Zhang M, Wing RA, Cregan PB, Rokhsar DS, Jackson SA (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46 (7):707–713

14. McGee R (2013) Mendel's legacy: an international pea sequencing project. Pisum Genet 44:13–14
15. Alves-Carvalho S, Aubert G, Carrère S, Cruaud C, Brochot A-L, Jacquin F, Klein A, Martin C, Boucherot K, Kreplak J, da Silva C, Moreau S, Gamas P, Wincker P, Gouzy J, Burstin J (2015) Full-length de novo assembly of RNA-seq data in pea (*Pisum sativum* L.) provides a gene expression atlas and gives insights into root nodulation in this species. Plant J 84:1–19
16. Pflieger S, Blanchet S, Meziadi C, Richard MM, Thareau V, Mary F, Mazoyer C, Geffroy V (2014) The "one-step" *Bean pod mottle virus* (BPMV)-derived vector is a functional genomics tool for efficient overexpression of heterologous protein, virus-induced gene silencing and genetic mapping of BPMV *R*-gene in common bean (*Phaseolus vulgaris* L.) BMC Plant Biol 14:232
17. Pflieger S, Blanchet S, Meziadi C, Richard MMS, Geffroy V (2015) *Bean pod mottle virus* (BPMV) viral inoculation procedure in common bean (*Phaseolus vulgaris* L.) Bio-protocol 5:e1524. http://www.bio-protocol.org/e1524
18. Meziadi C, Blanchet S, Richard MMS, Pilet-Nayel M-L, Geffroy V, Pflieger S (2016) *Bean pod mottle virus*: a new powerful tool for functional genomics studies in *Pisum sativum*. Plant Biotechnol J 14(8):1777–1787
19. Constantin GD, Krath BN, MacFarlane SA, Nicolaisen M, Johansen IE, Lund OS (2004) Virus-induced gene silencing as a tool for functional genomics in a legume species. Plant J 40:622–631

Chapter 24

Re-expressing Epigenetically Silenced Genes by Inducing DNA Demethylation Through Targeting of Ten-Eleven Translocation 2 to Any Given Genomic Locus

Julio Cesar Rendón, David Cano-Rodríguez, and Marianne G. Rots

Abstract

Epigenetic editing is a novel methodology to modify the epigenetic landscape of any genomic location. As such, the approach might reprogram expression profiles, without altering the DNA sequence. Epigenetic alterations, including promoter hypermethylation, are associated with an increasing number of human diseases. To exploit this situation, epigenetic editing rises as a new alternative to specifically demethylate abnormally hypermethylated regions. Here, we describe a methodology to actively demethylate the hypermethylated *ICAM-1* promoter. Reducing DNA methylation in our target region increased the expression of the *ICAM-1* gene. As the *ICAM-1* gene in our cell lines was highly methylated (up to 80%), this approach proves a robust manner to reduce methylation for hypermethylated regions. Epigenetic editing therefore not only provides an approach to address mechanisms of gene expression regulation, but also adds to the therapeutic toolbox as current inhibitors of epigenetic enzymes are limited by genome-wide effects.

Key words Epigenetic editing, *ICAM-1*, TET2, DNA demethylation, Zinc finger, Pyrosequencing, Transduction

1 Introduction

Epigenetics can be defined as the study of stable changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence [1]. Epigenetic changes include chemical groups (marks) covalently attached to DNA and/or histones, changing their properties and altering their function. One of the most intensively studied epigenetic marks is the cytosine methylation (5mC), which along with other less common modifications [2], occur generally in cytosines preceding a guanine (CpG). It consists of the covalent addition of a methyl group (CH_3) to the position 5 of the pyrimidine ring of the cytosine. This reaction is mediated by a family of DNA methyltransferase proteins (DNMTs) which either maintain the methylation status on

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_24,

newly formed DNA chains (DNMT1) or methylate previously unmethylated positions (de novo DNMT3A and DNMT3B).

CpG sites for DNA methylation can be found either clustered in highly CG dense regions (called CpG islands (CGI)), scattered in less condensed CpG site regions (CpG shores), or dispersed along the genome as independent CpG sites. Depending on the genomic location, CpG methylation functions to repress gene transcription [3]: when located within promoter regions (transcription factor inaccessibility), close to known transcription start sites (TSS) of genes (diminishing expression of the mRNA or inducing alternative TSS usage), or inside the gene bodies (repressing the activity of intragenic promoters—such as those driving the expression of non-coding RNAs—and even related with splicing alterations) [4–10]. DNA methylation regulates transcription by altering the molecular structure of the cytosines which impairs the interaction between DNA and their binding proteins, such as transcription factors, which often are sensible to this mark [11]. On the other hand, methylated CpG sites are recognized by methylated CpG binding proteins (like MBDs) [12] which recruit other transcription repressors as histone deacetylases (HDACs), inducing chromatin structure modifications which will also influence the transcription factors and transcription machinery accessibility.

As all epigenetic marks, DNA methylation is a dynamic and reversible process, and DNA demethylation can occur as an active or passive event. Passive DNA demethylation occurs when a DNA methylation mark is not copied to the newly formed DNA strand after replication; this creates hemimethylated sites, which will be lost upon subsequent cell division. Active DNA demethylation, on the other hand, refers to the enzymatic process by which 5mC mark is oxidized in several steps to 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxylcytosine (5mC → 5hmC → 5fC → 5caC). To recover the cytosine base in DNA chain, 5fC and 5caC can be excised by thymidine DNA glycosylase (TDG) leaving an abasic site which in turn is recognized and repaired with an unmodified cytosine by base excision repair (BER) system [13].

Human Ten-Eleven Translocation (TET) family enzymes, previously related to hematopoietic malignances, have been identified in 2009 as the responsible enzymes for 5mC → 5hmC oxidation; its identification as possible mediator in 5mC → 5hmC oxidation was initially identified due to their homology to the trypanosomal proteins JBP1 and JBP2 (members of the Fe(II)/α-ketoglutarate (α-KG)-dependent oxygenase family of enzymes) which mediate thymidine oxidation to 5-hydroxymethylurasil (5hmU) in this organism. Further functional experiments successfully validated TET family enzymes as responsible of 5mC oxidation in humans [14–16]. At the time of identification of the TET enzymes, we were pioneering the concept of epigenetic editing using engineered zinc finger proteins [17–20], through which a desired genomic region

can be targeted and any epigenetic mark present nearby can subsequently be modified by the fused epigenetic enzyme. This rewriting of the local epigenetic signature can potentially be further optimized to obtain the desired stable effect on gene expression.

Targeting specific regions of any genomic location nowadays can be achieved through different methodologies, most of them taking advantage of DNA binding domains (DBDs) of naturally occurring proteins, including helix-turn-helix, zinc fingers, leucine zippers, winged helix, helix-loop-helix, HMG-box, immunoglobulin fold, B3 domain, and more recently transcription activator like effectors (TALEs). These domains vary in size, length recognition capacity, affinity, and structure. Specific DNA sequence recognition can also be made using a small single stranded DNA sequence in what is called Triplex Forming Oligonucleotide (TFO), or by the combination of RNA-protein as the case of the CRISPR/CAS system [21–24].

Development of DNA targeting methodologies, especially the CRISPR-Cas revolution, currently allows us to regard the genome as a tunable structure, susceptible of editing. Most attention is being paid to specific induction of genome sequence changes by targeting DNA nucleases to a given genomic locus (genome editing). The introduced double strand break will either induce homologous recombination or is repaired by the error-prone nonhomologous end joining; the latter can be used to inactivate genes and this concept has been tested in various clinical trials using engineered ZFPs [25]. Additionally, the possibility to modulate gene expression without affecting the genome sequence by using epigenetic proteins, fused to these DNA targeting tools, is now increasingly appreciated. These fusion proteins can modify the epigenetic landscape of any genomic region (epigenetic editing) in order to induce or prevent gene expression.

In both genetic and epigenetic editing, programmable DNA targeting strategies are always necessary to interact with a desired region of DNA. The most common DNA targeting platforms for epigenetic editing include zinc fingers, TALEs, and the CRISPR/dCas9 system (targeting a catalytically dead Cas9 protein). The first and most extensively studied programmable DBDs are the zinc finger proteins, which are modular proteins consisting of individual "fingers" able to recognize three nucleotides each. Each finger is composed of around 30 amino acids, stabilized by a zinc ion which frequently binds to two cysteine and two histidine residues (Cys2His2-type). Based on its modular character, fusion of six fingers together extends the recognition sequence size to 18 base pairs which is enough to recognize unique sites in the human genome. Vast information about zinc fingers DNA binding rules resulted in the postulation of the "recognition code" [26–28], which allow the rational designing of zinc fingers targeting any sequence.

An increasing amount of human diseases including different kinds of cancer and syndromes [29] are now being related to epigenetic abnormalities such as aberrant histone modifications and hypo- or hyper-DNA methylation related with atypical expression of certain genes [30–32]. Specifically in cancer, transcriptional silencing by hypermethylation has been reported for key regulatory genes related with cell cycle or apoptosis; in this kind of situations, enforced DNA demethylation via targeting of such altered genes, using epigenetic editing, offers a novel approach for intervention and correction of specific gene expression abnormalities. Currently, DNA demethylation is clinically achieved with treatments as azacitidine (5-azacitidine) and decitabine (5-aza-2′deoxicitidine), but these treatments function in a genome-wide way and thus are limited in their clinical applications.

Several databases are available these days for a number of human and nonhuman cell lines and patient samples where gene expression is coupled to epigenetic marks. Such studies facilitate the choice of potential study models for epigenetic editing studies. In this regard, *C13ORF18* was identified as frequently hypermethylated in cervical cancer specimens, but not in normal cervix scraping [33]. We assigned tumor suppressive function to this gene by reactivating its expression, first by targeting the transcriptional activator VP64 [34], next by reexpression via induced DNA demethylation. The concept was further validated for other tumor suppressor genes [35, 36].

In our first report on inducing DNA demethylation, we compared TET1, TET2, and TET3 and concluded that TET3 was not effective and that TET2 was somewhat more effective than TET1 [37]. Our first model genes for targeted demethylation were the InterCellular Adhesion Molecule 1 (*ICAM-1*) and the Epithelial Cell Adhesion Molecule (*EpCAM*) gene promoter. The *ICAM-1* promoter is known to be silenced by specific CGI-related hypermethylation and previous reports demonstrate it is susceptible for reactivation using zinc finger protein coupled to a transient activation domain (VP64) [38, 39].

In this chapter, we present the protocol specially designed for lowering the methylation percentage of the hypermethylated *ICAM-1* promoter in human ovarian carcinoma cell line A2780. Here we induce zinc finger-TET2 fusion protein expression using a retroviral GFP reporter model; transduced cells are sorted from untransduced cells by FACS and finally *ICAM-1* expression and promoter methylation status is quantified by qRT-PCR and pyrosequencing, which allow confirmation of TET2 activity in the chosen model. Other DNA targeting platforms have further validated the potency of targeting TET1 to reduce hypermethylation status inducing gene expression [40, 41]. Altogether, the approach of targeted demethylation opens realistic avenues to start considering therapeutic reexpression of aberrantly silenced genes or of (fetal) genes which can compensate for a genetic mutation.

2 Materials

2.1 Transduction and Selection of ZF-TET2 Expressing Cells

2.1.1 Transduction of Cells to Express ZF-TET2

1. HEK293T and A2780 cell lines (human embryonic kidney and ovarian cancer cells, obtained from ATCC).
2. HEK293T cell culture media: 500 mL Dulbecco's Modified Eagle Medium (DMEM), 50% fetal bovine serum, 2 mM L-glutamine, 0.06 mg/mL gentamicin.
3. Calcium Chloride ($CaCl_2$) 2.5 M: 36.75 g $CaCl_2$*$2H_2O$, 100 mL H_2O miliQ (mQ). Filter the solution using a 0.2 μm filter.
4. Polybrene (Hexadimethrine bromide) solution 1 mg/mL: 50 mg Polybrene, 50 mL H_2O mQ.
5. PBS: 140 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, pH 7.3.
6. HBS 2× solution: 0.28 M NaCl, 50 mM HEPES, 1.42 mM Na_2HPO_4, pH 7.06. Filter the solution using a 0.2 μm filter.
7. VSV-G envelope expression plasmid (pMD2.G) and Gag-Pol expression plasmid (pMDLg/pRRE) [34].
8. Zinc finger expression vector (pMX-IRES-GFP-ZF-ED) [38].
9. 10 cm cell culture plates.
10. T75 flask.
11. 0.45 μm SFCA syringe filters.

2.1.2 Sorting of GFP Positive Cells

1. PBS four salts.
2. TEP (Trypsin-EDTA-PBS) solution: 487.5 mL PBS, 10 mL trypsin, 0.05 mM EDTA HEK293T cell culture media.

2.2 Quantification of DNA Methylation

2.2.1 DNA Extraction and Bisulfite Conversion

1. 6 well culture plates.
2. Chloroform.
3. NaCl 6 M solution.
4. Ice-cold Isopropanol.
5. Ice-cold Ethanol 75%.
6. EDTA 0.5 M, pH 8.
7. SDS 10%.
8. SE 10× Solution: 0.75 M NaCl, 0.25 M EDTA 0.5 M, pH 8.
9. TE buffer: 10 mM Tris and 0.1 mM EDTA, pH 8.
10. Proteinase K.
11. RNase.
12. EZ DNA methylation-Gold kit (ZYMO).

2.2.2 Pyrosequencing

1. Agarose.
2. Tris-Acetate-EDTA (TAE) Buffer: 40 mM Tris base, 2 mM EDTA, 20 mM acetic acid.
3. Streptavidin Sepharose High performance (GE Healthcare).
4. PyroMark™ Q24 pyrosequencer (*see* **Note 1**).
5. PyroMark™ Q24 vacuum work station.
6. PyroMark™ Q24 plates.
7. PyroMark™ PCR kit.
8. PyroMark™ Wash Buffer.
9. PyroMark™ Denaturation Buffer.
10. PyroMark™ Annealing Buffer.
11. PyroMark™ Binding Buffer.
12. PyroMark™ Gold Q24 reagents.
13. Primer Pyro-F (5′-GGGGAAGTTGGTAGTATTTAAAAGT-3′).
14. Primer Pyro-R (bio-5′-CCTTCCCCTCCCAAACAAATAC-TACAATTA-3′).
15. Primer Pyro-seq (5′-TGGGGGAGGGGAGTTTATT-3′).

3 Methods

3.1 Transduction and Selection of Cells to Express ZF-TET2

3.1.1 Zinc Finger Construction

1. Determine the sequence to be targeted (*see* **Note 2**).
2. ZF are constructed by building blocks, using previously described recognition modules as described by Barbas [42–44], Kim [45], or Young [46]. Nowadays, it is possible to use bioinformatics tools to select target sequences and design the proper ZF (*see* **Note 3**).
3. Assemble the ZF blocks using canonical peptide TGEKP as linker. We suggest ordering the complete ZF as an artificial minigen, avoiding manual assembly of different blocks.
   (Optional: Clone all different ZFs into proper bacterial expression plasmids; expressed ZF must be purified and dissociation constant (K_d) (*see* **Note 4**) can be determined by electromobility-shift assay (EMSA), gel shift assay, enzyme-linked immunosorbent assay (ELISA) surface plasmon resonance or a BIAcore system (*see* **Note 5**)).
4. Clone the selected ZF into a mammalian plasmid vector or viral (stable) expression system (*see* **Note 6**).

3.1.2 Transduction of Cells

1. Day −1: Seed 3.5–4 million HEK293T cells in ten dishes along with 10 mL of medium (three dishes per each plasmid to be transduced (in this example: nine plates, corresponding to three constructs) plus one additional dish as control).

2. Day 0: Refresh medium between 2 and 3 h before transfection using 5 mL of HEK293T medium.
3. Day 0: Prepare the transfection mixture: 200 μL mQ water, 7.5 μg plasmid mix (2.5 μg of pMD2.G plasmid, 5.0 μg of pMDLg/pRRE plasmid); fill up to 450 μL with mQ then add 50 μL $CaCl_2$ (2.5 M) drop wise and mix, leave at room temperature (prepare three times this mix, one per dish to be transduced).
4. Day 0: Into a tube containing 1500 μL of HBS 2× solution, add the complete (three times mixture) transfection mixture (*see* **Note 7**); incubate for 20–30 min, and add a final volume of 1 mL of the solution to each HEK293T cell dishes. Swirl the plates and put them at 37 °C 5% CO_2 incubator (*see* **Note 8**).
5. Day 1: Refresh the medium of treated HEK293T cell using 5 mL of prewarmed medium and seed host cells (A2780) in T75 flask using 700,000 cells in 15 mL of medium.
6. Day 2: Collect HEK293T cells supernatant medium in a single tube according to the plasmid to be transduced (±5 mL virus medium per dish, three dishes per plasmid), and add 5 mL of new medium to the each HEK293T cells dishes and put them back in the virus 37 °C CO_2 incubator.
7. Day 2: Centrifuge all collected virus medium to spin down the debris and cells (1000 × *g* for 10 min) and filter each supernatant trough a 0.45 μm SFCA filter.
8. Day 2: Per each HEK293 cell dish used previously (5 mL), and in independent tubes (one tube per construct to be transduced) 400 μmL FCS (final conc. 10%) and 30 μL polybrene (final conc. 6 μg/mL). Transfer the filtered virus containing media to the correct amount of solution (FCS + Polybrene) and mix gently.
9. Day 2: From the appropriate A2780 cell flask, aspirate the medium and add 7 mL of the corresponding virus medium (and save the remaining virus medium at 4 °C). Leave the cells at 37 °C incubator for 8 h, at the end of which, the virus medium of the A2780 cell flask must be replaced with previously saved virus medium. Finally put back the cells in the virus 37 °C CO_2.
10. Day 3: Repeat the procedure made on day 2 to the HEK293T cell supernatant and host A2780 cells. The virus producer cell HEK293T can be discarded after the second virus collection.
11. Day 4: Refresh medium of the host cells.

3.1.3 Sorting of Transduced GFP Positive Cells

Day 5+: In order to select the transduced cells, take advantage of the GFP protein expressed along with the zinc finger protein from the pMX-CD54-Opt31 plasmids (*see* **Note 9**).

Table 1
General cell culture plates culturing characteristics

Cell culture plate	Well diameter (mm)	Approx. growth area (cm^2)	Average cell seeding density	Working volume (mL)
6 well	35.8	9.5	3.0×10^5	1.9–2.9
12 well	22.1	3.8	1.0×10^5	0.76–1.14
24 well	15.6	1.9	0.5×10^5	0.38–0.57
96 well	11	0.95	1.3×10^4	0.19–0.28

1. Discard the medium and wash the cells three times with prewarmed PBS.
2. Add 500 μL TEP, spread around the flask, and incubate at 37 °C until cells detach (*see* **Note 10**).
3. Inactivate TEP by adding 2 mL of fresh medium and resuspend cells on it.
4. Collect the cells and centrifuge at 500 × *g* for 5 min.
5. Discard supernatant and resuspend the cell pellet in prewarmed PBS.
6. Proceed to FACS sorting (*see* **Note 11**).
7. Recover sorted cells in the proper medium and seed the cells in an appropriated multiple well plate according to the number of recovered cells (Table 1) (*see* **Note 12**).

3.2 Quantification of DNA Methylation

3.2.1 DNA Extraction and Bisulfite Conversion

1. Culture transduced cells in 6 well plates.
2. Discard the medium and wash the cells three times with prewarmed PBS.
3. Add 200 μL TEP, spread around the well, and incubate at 37 °C until cells detach.
4. Inactivate TEP and resuspend cells by adding 2 mL of fresh medium.
5. Collect the cells and centrifuge at 500 × *g* for 5 min.
6. Discard supernatant and save cell pellet at −80 °C or proceed directly to DNA extraction.
7. Incubate cell pellet at 55 °C for 5–10 min.
8. Resuspend pellet in 500 mL SE 1× solution.
9. Add 1 μL RNase and incubate for 1 h at RT.
10. Add 5 μL proteinase K and 50 μL SDS 10%, mix by inverting tube several times.
11. Incubate for at least 2 h at 55 °C mixing constantly by rotating upside down (*see* **Note 13**).
12. Add 222 μL (0.4 volume) NaCl 6 M, shake vigorously.

13. Add 777 μL (1 volume) Chloroform, shake the tube until two layers are completely mixed.
14. Rotating tubes upside down for at least 20 min up to 1 h.
15. Centrifuge samples at 13,400 × *g* 10 min at RT (*see* **Note 14**).
16. Carefully pipet out the upper layer and save in a clean 1.5 mL tube (*see* **Note 15**).
17. Measure the collected volume and add 1 volume of ice-cold Isopropanol, mix until white threads of DNA form a visible clump.
18. Centrifuge samples at >8000 × *g* 15 min at 4 °C.
19. Carefully pipette out the supernatant without disturbing the DNA pellet.
20. Add 500 μL of ice-cold ethanol 70% and resuspend the pellet.
21. Centrifuge >8000 × *g* 5 min at 4 °C, carefully decant the ethanol and leave the pellet to air dry at room temperature.
22. After all the ethanol is evaporated, add 30 μL TE buffer or mQ.
23. Quantify the DNA (e.g., by Nanodrop).
24. Storage: at 4 °C or at −20 °C for longer periods.
25. Use any of the commercially available bisulfite conversion kit to ensure high conversion ratio and as much as possible converted DNA recovery (*see* **Note 16**).
26. CpG sites in the converted DNA can be evaluated by Pyrosequencing or bisulfite sequencing. We suggest pyrosequencing to quantify the methylation % of each CpG site.

3.2.2 Pyrosequencing

1. Proceed to PCR amplification reaction using PyroMark™ PCR kit using protocol described by manufacturer (Table 2) (*see* **Notes 17** and **18**).

Table 2
PCR conditions for PyroMark™ PCR kit usage

Component	Final concentration	Cycling protocol		
PyroMark™ PCR Master Mix 2×	1×			
CoralLoad concentrate 10×	1×			
Q-Solution 5×	1×	94 °C	10 min	
Primer ICAM-1 pyro Fw 20 μM (*see* **Note 17**)	0.2 μM	94 °C	30 s	
Primer ICAM-1 pyro Rv 20 μM	0.2 μM	56 °C	30 s	45 cycles
RNase-free water	–	72 °C	30 s	
Template DNA BS 20 ng	1.6 ng/μL	72 °C	10 min	
Total volume (after adding template DNA)	25 μL			

2. Run up to 5 μL of PCR product in an agarose gel 1% and identify the positive amplification (*see* **Note 19**).
3. Pyrosequencing procedure is made using PyroMark™ Q24 pyrosequencer.
4. Prior to pyrosequencing procedure, use PyroMark™ Assay Design software to determine the position in the PyroMark™ Q24 plate and the amount of enzyme, substrate, and each dNTP needed for the assay and save the running protocol on a USB stick.
5. For each sample to be tested, dilute the sequencing primer to 0.3 μM in 25 μL.
6. Fill all the stations of the PyroMark™ Q24 vacuum workstation, with the appropriate volumes of Ethanol 70%, Denaturation solution, wash buffer, and water.
7. In 0.2 mL tubes, mix 20 μL of PCR product, with 1 μL of Streptavidin Sepharose™ High Performance beads, 40 μL of PyroMark™ Binding Buffer, and 19 μL of mQ water, shake the solution at ±1400 rpm up to 15 min using a horizontal shaker (e.g. Mikura Ltd).
8. Add to the previously defined wells (**step 4**) of the PyroMark™ Q24 plate, 25 μL of sequencing primer and place it in the PyroMark™ Q24 vacuum workstation.
9. Place the 0.2 mL tubes in the proper position with respect to the PyroMark™ Q24 plate in the PCR tube plate of the PyroMark™ Q24 vacuum workstation.
10. Using PyroMark Q24 vacuum tool, aspirate all beads, and transfer the vacuum tool throw the steps 1, 2, and 3 of the PyroMark™ Q24 vacuum workstation and release the beads into the PyroMark™ Q24 plate (*see* **Note 20**).
11. Incubate the PyroMark™ Q24 plate containing sequencing primers and streptavidin-PCR product beads at 80 °C for 2 min and let it cool down at RT.
12. Place the PyroMark™ Q24 plate in the PyroMark™ Q24 pyrosequencer.
13. Fill the PyroMark™ Q24 Cartridge with the proper amount of enzyme, substrate, and dNTPs previously determined in **step 4**.
14. Place the PyroMark™ Q24 Cartridge and Pyromark Q24 plate in the PyroMark™ Q24 pyrosequencer and run the protocol saved from **step 4**.
15. After running, determine the percentage of methylation in each step using PyroMark Assay Design software.

4 Notes

1. We use PyroMark™ pyrosequencing system from QIAGEN.
2. This is the key step for zinc finger design, DNase I sensitive regions might provide good targets for ZF binding, although for some DNase I nonsensitive sites effective ZFs have been designed. We advise to design and screen a panel of ZFs for the same region [47], as some might work in one cell line, while others are active in other lines (*see*, for example, Huisman et al. [34]).
3. The target site is divided in N pieces of 3 bp or overlapping 4 bps segments where the last base of the 4 bp block is the first base of the next block. Modules for ZF sequence design can be found online (http://www.scripps.edu/barbas/zfdesign/zfdesignhome.php) or via other references [42, 46, 43–45, 48].
4. Specificity of the designed ZF can be examined by testing mutated target sites.
5. Several systems for ZF binding properties can be used as systematic evolution of ligands by exponential enrichment (SELEX), cyclic amplification and selection of targets (CAST), DNA microarrays or even bacterial one-hybrid system [49–51].
6. Zinc finger targeting efficiency could be verified by transient reporter assay or by measuring expression levels of the target gene.
7. Add the transfection mixture drop wise while, at the same time, using a mechanical pipettor and a 1 mL pipet directly at the bottom, blow small bubbles into the solution to favor the dilution.

   Use a mechanical pipettor attached to a plugged 1- or 2-mL pipet to bubble the HBS and add the DNA/$CaCl_2$ mixture.
8. Because of the infectious potential of the virions being produced, special cautions should be taken regarding laboratory space chosen for this procedure.
9. Usually low cytotoxicity is observed and 2 days after transduction, infected cells can be selected.
10. A2780 cells can take some time to detach completely; be sure to wash the cells properly and to cover the whole flask with TEP, even adding up to 1 mL TEP.
11. Be aware to resuspend completely the cells to avoid needle obstructions, use the proper needle according to the size of your cells. Use the negative control cells to calibrate the Fluorescence activated cell sorter.

12. Make passages of cells while increasing the growing surface until enough cells are available for DNA extraction and analysis (and we suggest to also freeze some ampoules).

    We suggest that all transduced cells with each construct are kept in culture for the same period of time at the moment of DNA extraction, in order to ensure direct comparison of the effect among the different zinc finger constructs.

13. Overnight incubation at 55 °C with constant mixing will increase protease action facilitating step 16 procedures.
14. Two colorless layers separated by a third white protein layer are observed. Avoid disturbing the layers, if it happens, recentrifuge 20 min and continue extraction.
15. Uncomplete protease reaction will leave protein fractions in the upper layer, usually connected to the intermediate layer, which during pipetting will drag the protein layer increasing contamination of final DNA.
16. We find that for the EZ DNA methylation-Gold kit, using 500 ng of gDNA and alternative protocol #2 of incubation 98 °C 10′ followed by 53 °C 4 h, good conversion ratios were obtained.
17. Test different cycling protocols (focusing on annealing step) in order to improve PCR amplification. Concentration can be tested and adapted depending on the amplification results; Q solution can also be obviated.
18. One of the primers must be biotinylated in order to perform pyrosequencing, in many cases an extended non-biotinylated primer Fw or Rv can be used; in this case, an additional universal biotinylated reverse primer must be used. This additional primer is complementary to the extended primer. We suggest a different cycling protocol is necessary for this conditions (Table 3).
19. PCR products for pyrosequencing can be stored up to 24 h at 4 °C, although we suggest using them as soon as possible due to observed decreased pyrosequencing efficiency in longer storages PCR products.
20. Special care must be taken when aspirate the bead, to ensure complete collection from tubes and when releasing of the beads exactly on the correct well of the PyroMark™ Q24 plate. We suggest keeping suspended the vacuum tool in the exact position before turning off the vacuum source.

Table 3
PCR conditions for PyroMark™ PCR kit usage using not biotinylated primersq

Component	Volume per reaction	Cycling protocol		
Step 1				
PyroMark™ PCR Master Mix 2×	1×			
CoralLoad Concentrate 10×	1×			
Q-Solution 5×	1×	94 °C	10 min	
Primer pyro Fw 20 μM	0.2 μM	94 °C	30 s	
Primer pyro Rv1 20μM[a]	0.04 μM	50°C[b]	30 s	20 cycles
RNase-free water	–	72 °C	30 s	
Template DNA BS 20 ng	1.6 ng/μL			
Total volume (after adding template DNA)	25 μL			

Step 2: Immediately add the biotinylated pyro rv2 primer[c] to a final concentration 0.16 μM and continue running for additional 20 cycles

[a]Extended sequence specific primer *not biotinylated*

[b]Melting temperature between first and second step must be adapted

[c]Biotinylated Rv primer targeting the extension of pyro rv1

References

1. Wu C, Morris JR (2001) Genes, genetics, and epigenetics: a correspondence. Science 293:1103–1105
2. Fu Y, He C (2012) Nucleic acid modifications with epigenetic significance. Curr Opin Chem Biol 16:516–524
3. Suzuki Y, Korlach J, Turner SW, Tsukahara T, Taniguchi J, Qu W, Ichikawa K et al (2016) AgIn: measuring the landscape of CpG methylation of individual repetitive elements. Bioinformatics 32:2911–2919
4. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA et al (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38:626–635
5. Kornblihtt AR (2006) Chromatin, transcript elongation and alternative splicing. Nat Struct Mol Biol 13:5–7
6. Aguilar JL, Montes A, Montero A, Vidal F, F-Llamazares J, Pastor C (1992) Continuous pleural infusion of bupivacaine offers better postoperative pain relief than does bolus administration. Reg Anesth 17:12–14
7. Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A 103:1412–1417
8. De Smet C, Lurquin C, Lethe B, Martelange V, Boon T (1999) DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter. Mol Cell Biol 19:7327–7335
9. Edwards CA, Ferguson-Smith AC (2007) Mechanisms regulating imprinted genes in clusters. Curr Opin Cell Biol 19:281–289
10. Shen L, Kondo Y, Guo Y, Zhang J, Zhang L, Ahmed S, Shu J et al (2007) Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. PLoS Genet 3:2023–2036
11. Buck-Koehntop BA, Defossez PA (2013) On how mammalian transcription factors recognize methylated DNA. Epigenetics 8:131–137
12. Boyes J, Bird A (1991) DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. Cell 64:1123–1134
13. He YF, Li BZ, Li Z, Liu P, Wang Y, Tang Q, Ding J et al (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science 333:1303–1307
14. Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, He C et al (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science 333:1300–1303

15. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S et al (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 324:930–935
16. Ito S, D'Alessio AC, Taranova OV, Hong K, Sowers LC, Zhang Y (2010) Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature 466:1129–1133
17. Siddique AN, Nunna S, Rajavelu A, Zhang Y, Jurkowska RZ, Reinhardt R, Rots MG et al (2013) Targeted methylation and gene silencing of VEGF-A in human cells by using a designed Dnmt3a-Dnmt3L single-chain fusion protein with increased DNA methylation activity. J Mol Biol 425:479–491
18. Rivenbark AG, Stolzenburg S, Beltran AS, Yuan X, Rots MG, Strahl BD, Blancafort P (2012) Epigenetic reprogramming of cancer cells via targeted DNA methylation. Epigenetics 7:350–360
19. Falahi F, Huisman C, Kazemier HG, van der Vlies P, Kok K, Hospers GA, Rots MG (2013) Towards sustained silencing of HER2/neu in cancer by epigenetic editing. Mol Cancer Res 11:1029–1039
20. de Groote ML, Verschure PJ, Rots MG (2012) Epigenetic editing: targeted rewriting of epigenetic marks to modulate expression of selected target genes. Nucleic Acids Res 40:10596–10613
21. Uil TG, Haisma HJ, Rots MG (2003) Therapeutic modulation of endogenous gene function by agents with designed DNA-sequence specificities. Nucleic Acids Res 31:6064–6078
22. Thakore PI, Black JB, Hilton IB, Gersbach CA (2016) Editing the epigenome: technologies for programmable transcription and epigenetic modulation. Nat Methods 13:127–137
23. Tost J (2016) Engineering of the epigenome: synthetic biology to define functional causality and develop innovative therapies. Epigenomics 8:153–156
24. Cano-Rodriguez D, Rots MG (2016) Epigenetic editing: on the verge of reprogramming gene expression at will. Curr Genet Med Rep 4 (4):170–179
25. Sangamo_BioSciences (2016) Sangamo BioSciences, Inc. Therapeutic applications. Hane Chow Inc. http://www.sangamo.com/technology/therapeutic-applications.html
26. Desjarlais JR, Berg JM (1992) Toward rules relating zinc finger protein sequences and DNA binding site preferences. Proc Natl Acad Sci U S A 89:7345–7349
27. Desjarlais JR, Berg JM (1993) Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. Proc Natl Acad Sci U S A 90:2256–2260
28. Choo Y, Klug A (1994) Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. Proc Natl Acad Sci U S A 91:11163–11167
29. Egger G, Liang G, Aparicio A, Jones PA (2004) Epigenetics in human disease and prospects for epigenetic therapy. Nature 429:457–463
30. Zhang X, Zhang J, Wang R, Guo S, Zhang H, Ma Y, Liu Q et al (2016) Hypermethylation reduces the expression of PNPLA7 in hepatocellular carcinoma. Oncol Lett 12:670–674
31. Manoochehri M, Borhani N, Karbasi A, Koochaki A, Kazemi B (2016) Promoter hypermethylation and downregulation of the FAS gene may be involved in colorectal carcinogenesis. Oncol Lett 12:285–290
32. Robertson KD (2005) DNA methylation and human disease. Nat Rev Genet 6:597–610
33. Yang N, Eijsink JJ, Lendvai A, Volders HH, Klip H, Buikema HJ, van Hemel BM et al (2009) Methylation markers for CCNA1 and C13ORF18 are strongly associated with high-grade cervical intraepithelial neoplasia and cervical cancer in cervical scrapings. Cancer Epidemiol Biomark Prev 18:3000–3007
34. Huisman C, Wisman GB, Kazemier HG, van Vugt MA, van der Zee AG, Schuuring E, Rots MG (2013) Functional validation of putative tumor suppressor gene C13ORF18 in cervical cancer by Artificial Transcription Factors. Mol Oncol 7:669–679
35. Huisman C, van der Wijst MG, Falahi F, Overkamp J, Karsten G, Terpstra MM, Kok K et al (2015) Prolonged re-expression of the hypermethylated gene EPB41L3 using artificial transcription factors and epigenetic drugs. Epigenetics 10:384–396
36. Huisman C, van der Wijst MG, Schokker M, Blancafort P, Terpstra MM, Kok K, van der Zee AG et al (2016) Re-expression of selected epigenetically silenced candidate tumor suppressor genes in cervical cancer by TET2-directed demethylation. Mol Ther 24:536–547
37. Chen H, Kazemier HG, de Groote ML, Ruiters MH, Xu GL, Rots MG (2014) Induced DNA demethylation by targeting Ten-Eleven Translocation 2 to the human ICAM-1 promoter. Nucleic Acids Res 42:1563–1574
38. Magnenat L, Blancafort P, Barbas CF 3rd (2004) In vivo selection of combinatorial libraries and designed affinity maturation of polydactyl zinc finger transcription factors for ICAM-1 provides new insights into gene regulation. J Mol Biol 341:635–649

39. de Groote ML, Kazemier HG, Huisman C, van der Gun BT, Faas MM, Rots MG (2014) Upregulation of endogenous ICAM-1 reduces ovarian cancer cell growth in the absence of immune cells. Int J Cancer 134:280–290
40. Maeder ML, Angstman JF, Richardson ME, Linder SJ, Cascio VM, Tsai SQ, Ho QH et al (2013) Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. Nat Biotechnol 31:1137–1142
41. Stolzenburg, S., Goubert, D., Rots, M. G. (2016) Rewriting DNA methylation signatures at will: the curable genome within reach? In: Jeltsch A., Jurkowska R. DNA methylation. Springer International Publishing, Switzerland. Adv Exp Med Biol 945:475–490
42. Segal DJ, Dreier B, Beerli RR, Barbas CF 3rd (1999) Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5′-GNN-3′ DNA target sequences. Proc Natl Acad Sci U S A 96:2758–2763
43. Dreier B, Beerli RR, Segal DJ, Flippin JD, Barbas CF 3rd (2001) Development of zinc finger domains for recognition of the 5′-ANN-3′ family of DNA sequences and their use in the construction of artificial transcription factors. J Biol Chem 276:29466–29478
44. Dreier B, Fuller RP, Segal DJ, Lund CV, Blancafort P, Huber A, Koksch B et al (2005) Development of zinc finger domains for recognition of the 5′-CNN-3′ family DNA sequences and their use in the construction of artificial transcription factors. J Biol Chem 280:35588–35597
45. Bae KH, Kwon YD, Shin HC, Hwang MS, Ryu EH, Park KS, Yang HY et al (2003) Human zinc fingers as building blocks in the construction of artificial transcription factors. Nat Biotechnol 21:275–280
46. Maeder ML, Thibodeau-Beganny S, Osiak A, Wright DA, Anthony RM, Eichtinger M, Jiang T et al (2008) Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. Mol Cell 31:294–301
47. Liu PQ, Rebar EJ, Zhang L, Liu Q, Jamieson AC, Liang Y, Qi H et al (2001) Regulation of an endogenous locus using a panel of designed zinc finger proteins targeted to accessible chromatin regions. Activation of vascular endothelial growth factor A. J Biol Chem 276:11323–11334
48. Mandell JG, Barbas CF 3rd (2006) Zinc Finger Tools: custom DNA-binding domains for transcription factors and nucleases. Nucleic Acids Res 34:W516–W523
49. Corbi N, Perez M, Maione R, Passananti C (1997) Synthesis of a new zinc finger peptide; comparison of its 'code' deduced and 'CASTing' derived binding sites. FEBS Lett 417:71–74
50. Bulyk ML, Huang X, Choo Y, Church GM (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proc Natl Acad Sci U S A 98:7158–7163
51. Meng X, Brodsky MH, Wolfe SA (2005) A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat Biotechnol 23:988–994

Chapter 25

Knockdown of Rice microRNA166 by Short Tandem Target Mimic (STTM)

Sachin Teotia, Dabing Zhang, and Guiliang Tang

Abstract

Small RNAs, including microRNAs (miRNAs), are abundant in plants and play key roles in controlling plant development and physiology. miRNAs regulate the expression of the target genes involved in key plant processes. Due to functional redundancy among miRNA family members in plants, an ideal approach to silence the expression of all members simultaneously, for their functional characterization, is desirable. Target mimic (TM) was the first approach to achieve this goal. Short tandem target mimic (STTM) is a potent approach complementing TM for silencing miRNAs in plants. STTMs have been successfully used in dicots to block miRNA functions. Here, we describe in detail the protocol for designing STTM construct to block miRNA functions in rice. Such approach can be applied to silence miRNAs in other monocots as well.

Key words miRNA, miR166, Rice, Short tandem target mimic (STTM), Target mimic

1 Introduction

Small RNAs, including microRNAs (miRNAs) and siRNAs, are implicated in controlling various plant functions. miRNAs control the expression of target genes by binding to the complementary sites in those mRNAs, which leads to their cleavage and/or translational blockage. Various approaches have been followed to study functions of miRNAs. These include either up- or downregulating the expression of target miRNAs. Knockdown of expression of all members of a miRNA family is desirable in order to study their loss-of-functions. This can be achieved by creation of target mimics [1], molecular sponges [2], and short tandem target mimics (STTMs) [3]. STTMs have been effective in knocking down miRNA expression by degrading them, and have been reported to knockdown target miRNA expression in Arabidopsis [3], tobacco [4], soybean [5], tomato [6], cotton [7], and wheat [8].

STTM is an artificial short (~100 nt) noncoding RNA that can be expressed either through stable plant transformation or through virus-based transient expression systems [4]. STTM consists of two

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_25,

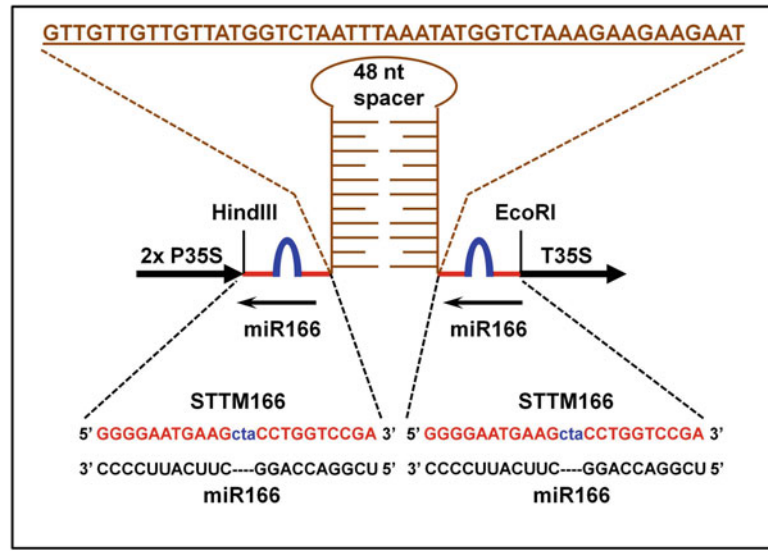


Fig. 1 STTM166 construct with miR166 binding sites (in *red*) on flanking sides separated by a 48-nt spacer (in *brown*) forming an imperfect weak stem-loop. The tri-nucleotide bulge is shown in *blue*. The complementary miR166 is shown in *black. 2x P35S, enhanced 35S promoter; T35S, terminator*

miRNA complementary binding sites separated by a 48–88 nt spacer. The binding sites have three nucleotide mismatches at the putative miRNA cleavage site. These nucleotide mismatches form a bulge which enables STTMs to escape the cleavage by the target miRNAs. This subsequently helps STTMs to sequester and/or degrade target miRNAs (Fig. 1). The spacer forms a weak secondary structure and not only helps in stabilizing the overall STTM structure but also helps in separating two miRNA binding sites from RISC collision [9]. To date, STTMs have not been reported to work in rice. Here, we describe the protocol of designing STTMs with an example of targeting miR166 in rice. This protocol can be applied to silence most miRNAs and other small RNAs, like siRNAs, in rice and other monocots.

2 Materials

2.1 Plant Growth Conditions

Rice (*Oryza sativa*) is grown in tissue culture media supplemented with various nutrients and plant hormones at 25–28 °C, under light.

2.2 Rice Transformation

1. AAM-AS Buffer: For 1 L AAM take 100 ml 10× AA macroelement, 10 ml 100× AA microelement I, 1 ml 1000× AA microelement II, 100 ml 10× AA, 5 ml 200× Fe, 10 ml 100× MS vitamin, 2.94 g KCl, 0.5 g casein acid hydrolysate, 68.5 g sucrose, 36 g glucose, 0.1 g inositol.

Adjust pH to 5.2, sterilize (121 °C for 20 min). Add acetosyringone (200 μM/L in DMSO) before use (*see* **Note 1**).

10× AA macroelement (per L): 1.7 g KH_2PO_4, 3.7 g $MgSO_4{\cdot}7H_2O$, 4.4 g $CaCl_2{\cdot}2H_2O$

100× AA microelement I (per L): 1.69 g $MnSO_4{\cdot}H_2O$, 0.86 g $ZnSO_4{\cdot}7H_2O$, 0.62 g H_3BO_3, 0.083 g KI.

1000× AA microelement II (per L): 0.025 g $CuSO_4{\cdot}5H_2O$, 0.25 g NaMoO4·$2H_2O$, 0.025 g $CoCl_2{\cdot}6H_2O$.

10× AA (per L): 8.77 g glutamine, 2.66 g aspartic acid, 2.88 g arginine, 0.75 g glycine.

200× Fe (per 500 ml): 3.73 g Na_2-EDTA, 2.78 g $FeSO_4{\cdot}7H_2O$.

100× MS vitamin (per L): 0.05 g nicotinic acid, 0.1 g aneurine hydrochloride, 0.05 g vitamin B6.

2. NBD2 buffer: To make 1 L of NBD2 take 50 ml 20× macroelement I, 25 ml 40× macroelement II, 25 ml 40× macroelement III, 10 ml 100× B5 microelement I, 1 ml 1000× B5 microelement II, 10 ml 100× B5 vitamin I, 10 ml 100× B5 vitamin II, 5 ml 200× Fe, 30 g sucrose, 0.1 g inositol, 0.5 g L-proline, 0.5 g L-glutamine, 0.5 g casein acid hydrolysate. Adjust pH to 5.8, sterilize (121 °C for 20 min).

   20× macroelement I (per L): 56.6 g KNO_3, 9.26 g $(NH_4)_2SO_4$, 8 g KH_2PO_4.

   40× macroelement II (per 500 ml): 3.32 g $CaCl_2{\cdot}2H_2O$.

   40× macroelement III (per 500 ml): 3.7 g $MgSO_4{\cdot}7H_2O$.

   100× B5 microelement I (per L): Solution A—Add 0.781 g $MnSO_4{\cdot}H_2O$ and 0.2 g $ZnSO_4{\cdot}7H_2O$ in 400 ml H_2O. Solution B—Add 0.3 g H_3BO_3 and 0.075 g KI in 400 ml H_2O. Slowly mix solution A and B, add H_2O to make 1 L.

   1000× B5 microelement II (per 500 ml): 0.125 g $Na_2MoO_4{\cdot}2H_2O$, 0.0125 g $CuSO_4{\cdot}5H_2O$, 0.0125 g $CoCl_2{\cdot}6H_2O$.

   100× B5 vitamin I (per 500 ml): 0.5 g Aneurine hydrochloride, 0.05 g vitamin B6, 0.05 g nicotinic acid.

   100× B5 vitamin II (per 500 ml): 0.1 g Glycine.

   100× 2,4-D (2,4-Dichlorophenoxyacetic Acid) (per L): 0.2 g 2,4D.

   200× Fe: (per 500 ml): 3.73 g Na_2-EDTA, 2.78 g $FeSO_4{\cdot}7H_2O$.

3. NBD-AS buffer: To make 1 L NBD-AS take the mother solution of NBD2, and add 1 g casein acid hydrolysate, 0.1 g inositol, 30 g sucrose, 10 g glucose. Adjust pH to 5.2, sterilize (121 °C for 20 min).

   Add acetosyringone (100 μM/L) before use.

4. Regeneration medium (per L):

   4.4 g Murashige and Skoog basal medium w/vitamins (MS), 0.5 g casein acid hydrolysate, 2 mg 6-BA, 0.5 mg KT,

0.5 mg NAA, 30 g sucrose, 15 g D-sorbitol. Adjust pH to 5.8, sterilize (121 °C for 20 min).

5. Rooting medium (per L):
   2.2 g MS medium, 30 g sucrose. Adjust pH to 5.8, sterilize (121 °C for 20 min).
6. Antibiotics: hygromycin, carbenicillin, cefotaxime and timentin.
7. Ethanol, bleach (Clorox), Tween-20, Whatman filter papers #1, Petri plates, magenta box.

2.3 Construction of Recombinant pOT2 and pCAMBIA1301 Vectors

1. Vectors to be used: pOT2, pCAMBIA1301-PacI (Fig. 2).
2. Luria Bertani (LB) broth.
3. 30 mg/ml Chloramphenicol.
4. 50 mg/ml Kanamycin.

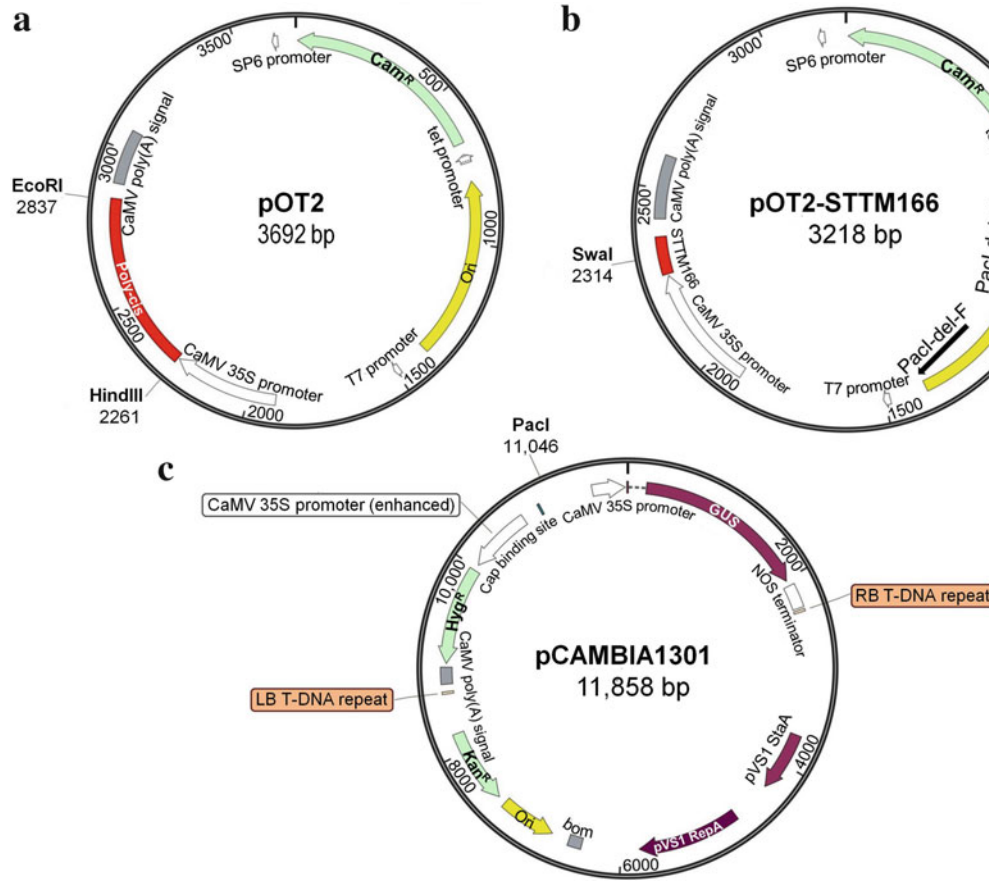


Fig. 2 Vectors used to silence miR166. (**a**) pOT2-poly-cis vector; (**b**) pOT2 vector with poly-cis site replaced with STTM166 construct (in *red*) and origin-PacI-del primers to delete replication origin region and create PacI site; (**c**) pCAMBIA1301 vector with PacI restriction site. Selectable markers: Cam^R, chloramphenicol; Hyg^R, hygromycin; Kan^R, kanamycin. Ori, origin of replication

Table 1
Primers used for production of STTM166–48 using pOT2-Poly-Cis template

STTM166-Ultra-*Hin*dIII-F (5′–3′)	agctGGGGAATGAAGctaCCTGGTCCGAgttgttgttgttatggtc taatttaaatatggtctaaagaagaagaatGGGGAATGAAGctaCCTG GTCCGA
STTM166-Ultra-*Eco*RI-R (5′–3′)	aattTCGGACCAGGtagCTTCATTCCCCattcttcttctttagaccat ATTTAAATtagaccataacaacaacaacTCGGACCAGGtagCTTC ATCCCCC

Table 2
Primers used for origin deletion in pOT2-STTM construct

Origin-del-PacI-PF (5′–3′)	TCCCTTAATTAAGTTTGCAAGCAGCAGATTACGCG
Origin-del-PacI-PR (5′–3′)	TCCCTTAATTAAGAAAGGCGGACAGGTATCCGGTAAG

5. 40 mg/ml Rifampicin.
6. Transformation competent *Escherichia coli* (DH5α).
7. Transformation competent *Agrobacterium tumefaciens* (EHA105).
8. Plasmid miniprep kit.
9. Restriction enzymes: PacI, SwaI, *Hin*dIII, and *Eco*RI.
10. Reaction buffers for restriction digestion.
11. STTM construction primers (Table 1), the origin deletion primers (Table 2).
12. *Taq* DNA polymerase.
13. Gel and PCR Clean-Up Kit.
14. Sephadex G-25 columns.
15. Agarose.
16. Gel electrophoresis Units.
17. T4 DNA ligase.
18. Laminar hood.
19. Incubators and shakers for growing *E. coli* and Agrobacterium.

2.4 RNA Extraction, First-Strand cDNA Synthesis, and Real-Time PCR Analysis

1. TRIzol reagent.
2. High-Capacity cDNA Reverse Transcription Kit.
3. 40 U/μl RNase inhibitor.
4. DNase-/RNase-free water.
5. Oligo(dT18) primer.
6. Sequence-specific primers for real-time PCR of target genes.
7. Real-time PCR kit, SYBR® Select Master Mix.

8. Nanodrop.
9. PCR primers for rice EF1α gene as endogenous control:
   Forward: 5′-CGCTCTTCTTGCTTTCACTCTTG-3′.
   Reverse: 5′-TAGGATGAGACTTCCTTCACGATTTC -3′.
10. miR166 stem-loop real-time primers:
   miR166-stemloop-RT-primer
   5′-GTTGGCTCTGGTGCAGGGTCCGAGGTATTCGCACCAGAGCCAACGGGGAA-3′
   universal reverse primer
   5′-GTGCAGGGTCCGAGGT-3′
   miR166-F
   5′-GTTTTCGGACCAGGCTTCA-3′.
11. Real-time PCR cycler.
12. Plant DNA isolation kit.

3 Methods

3.1 Designing STTM Construct

STTM structure targeting miR166 should be designed as follows: The mature miR166 sequence is 5′-UCGGACCAGGCUUCAUUCCCC-3′. The sequence for miR166 binding site should be complementary to it, which is 5′-GGGGAAUGAAGCCUGGUCCGA-3′.

In order for this site to bind the mature miR166 without being cleaved by it, a tri-nucleotide bulge must be designed in the cleavage region, corresponding between the 10th and 11th positions of the mature miR166. In this case, we have taken the tri-nucleotide sequences (CUA) from the Arabidopsis IPS1 that corresponds to 10th and 11th positions of miR399 [1] (*see* **Note 2**). After being inserted with "cua," the binding sequence becomes 5′-GGGGAAUGAAGcuaCCUGGUCCGA-3′. This sequence on being converted into DNA becomes 5′-GGGGAATGAAGctaCCTGGTCCGA-3′. Finally, the two binding sites separated by 48-nt spacer will form a STTM fragment, which will look like this:

5′-GGGGAATGAAGctaCCTGGTCCGAgttgttgttgttatggtctaatttaaatatggtctaaagaagaagaatGGGGAATGAAGctaCCTGGTCCGA-3′ (Fig. 1).

The above fragment with a restriction overhang of *Hind*III at the 5′ site was synthesized. Its complementary sequence was synthesized with *Eco*RI overhang at the 5′ site (Table 1).

3.2 Cloning STTM Fragment in pOT2-Poly-Cis Vector

The creation of pOT2-Poly-Cis vector has been described previously [10]. Here we describe how to clone STTM fragment first into pOT2-Poly-Cis vector and then sub-clone it into pCAMBIA-1301 binary vector.

1. Digest 1 μg of pOT2-Poly-Cis vector (Fig. 2a) with *Hin*dIII and *Eco*RI in reaction buffer at 37 °C for 4 h. This digestion will release poly-cis fragment from the pOT2 vector.
2. Run the digested product in 1% agarose gel and cut the upper band of about ~3 kb.
3. Elute the DNA using Gel and PCR Clean-Up Kit.
4. Mix 5 μl each of the two single-stranded synthesized STTM oligos from Table 1 and heat them at 95 °C for 5 min. Then allow them to cool gradually at room temperature. After cooling, the two fragments will hybridize with each other and form a double-stranded structure with restricted *Hin*dIII and *Eco*RI site overhangs at the 5′ and 3′ sites, respectively (*see* **Note 3**).
5. Take 10 μl of the above mix and ligate with 200 ng of the above digested and eluted DNA of pOT2-Poly-Cis vector in 20–25 μl reaction using 1 μl T4 DNA ligase and ligase buffer. Keep the above reaction overnight at 16 °C.
6. Transform the above reaction into chemically competent *Escherichia coli* (DH5α) cells by heat shock at 42 °C.
7. Select the bacterial cells on LB medium with 30 mg/ml chloramphenicol for overnight at 37 °C.
8. Take a few colonies and grow in overnight culture by shaking at 37 °C, isolate plasmid using plasmid isolation kit and digest by SwaI. Positive clones bearing STTM fragment will get linearized to show a band of about ~3.2 kb, as they have a SwaI site in the spacer region of STTM. The original pOT2-Poly-Cis vector has no SwaI site.
9. Verify the positive pOT2-STTM166 plasmid by DNA sequencing using the sequencing primers: STTM-common-real-PF (5′-catttggagaggacagcccaag-3′) and STTM-common-real-PR (5′-ctggtgatttcagcgtaccgaa-3′). Use the construct with correct STTM sequence in pOT2 vector for further experiments.
10. Take the DNA of the above positive clone and amplify it by PCR using the Origin-PacI-del primers (Table 2) (Fig. 2b) using *Taq* DNA polymerase. Use the cycling conditions as: 94 °C, 2 min; [94 °C, 30 s; 58 °C, 30 s; 68 °C, 4 min (30 cycles)]; 68 °C, 10 min (*see* **Note 4**).
11. After this PCR the PacI sites are added at each end of the PCR fragment (2846 bp long) and the "Replication origin" of the pOT2 vector is deleted.
12. Verify the PCR on a 1% agarose gel. Purify the PCR product using Sephadex G-25 column.

3.3 Sub-Cloning STTM Fragment from pOT2-Poly-Cis Vector into Binary Vector pCAMBIA-1301-PacI

The structure of STTM166, *Cauliflower mosaic virus* (CaMV) d35S (2× enhanced) promoter, the 35S terminator and chloramphenicol resistance (Cam^R) selection marker is subcloned into a binary vector pCAMBIA1301 for Agrobacterium-mediated transformation (*see* **Note 5**). Vector pCAMBIA1301 has a unique PacI site (Fig. 2c).

1. Mix 400 ng of PCR product in **step 12** of previous section with 400 ng of pCAMBIA1301 and digest them with 1 μl PacI in digestion buffer 1 in 50 μl reaction volume and incubate for 4 h at 37 °C. After digestion, heat inactivate the reaction mix at 65 °C for 20 min.
2. Purify the digested products using Sephadex G-25 column.
3. Take 26 μl of the above purified PacI-digested origin-deleted PCR product and pCAMBIA1301 vector and add 1 μl T4 DNA ligase and 3 μl of ligation buffer to make 30 μl reaction volume. Incubate the reaction mix and ligate at 16 °C overnight (*see* **Note 6**).
4. Transform the ligation reaction into *E. coli* and plate on a LB Agar plate with both kanamycin (50 mg/ml) and chloramphenicol (30 mg/ml) for screening the colonies containing the recombinant pCAMBIA1301-STTM construct. Kanamycin resistance (Kan^R) comes from the pCAMBIA1301 backbone and Cam^R comes from origin-deleted PCR product of pOT2 backbone. The double selection ensures all correct colonies, removing the possibility of false positive colonies.
5. Verify the positive construct by digesting with PacI enzyme. After digestion, correct construct will give two DNA bands on Agarose gel (one of about ~12 kb and the other is about ~3 kb). The positive construct is further verified by DNA sequencing.
6. Plasmid DNA of positive construct is isolated by a miniprep kit.

3.4 Mobilization of the Recombinant pCAMBIA1301-STTM166 Vector into A. tumefaciens by Chemical Transformation

The recombinant pCAMBIA1301-STTM166 vector is transformed into *Agrobacterium* strain EHA105 by freeze-thaw method [11].

1. Add 1 μg plasmid DNA dissolved in sterile water to a vial of transformation-competent *Agrobacterium* cells thawed in ice.
2. After transformation by freeze-thaw method, plate the transformed Agrobacterium cells on LB plates supplemented with 40 mg/ml rifampicin, 50 mg/ml kanamycin, and 30 mg/ml chloramphenicol.

3. Incubate the plates at 28 °C for 2–3 days until the colonies become bigger.
4. For screening the transformants, perform colony PCR using previously described STTM-common-real primers.
5. Prepare a glycerol stock of the confirmed Agrobacterium transformant and store at −80 °C for further use.

3.5 Transformation of Rice (Oryza sativa ssp. japonica cv. Nipponbare) with the Agrobacterium Harboring pCAMBIA1301-STTM166 Vector

Transform rice with the above construct and a control transformation with empty pCAMBIA1301 vector, as described previously [12] with slight modifications.

1. Take about 100 de-husked rice seeds and sterilize with 70% ethanol for 1 min followed by 30 min sterilization with 50% Clorox with 0.1% Tween-20.
2. Wash the sterilized seeds ten times with sterile double-distilled water in laminar hood.
3. Dry the seeds on sterile filter paper and culture them on NBD2 medium for 15 days at 25–28 °C in dark. Change the medium and keep for another 10 days on NBD2 medium in dark.
4. Inoculate a single colony each of Agrobacterium harboring pCAMBIA1301-STTM166 vector and empty pCAMBIA1301 vector in 5 ml liquid LB medium containing selective antibiotics (kanamycin, rifampicin, chloramphenicol), in a 50 ml conical sterile test tube. Shake on an orbital shaker at 250 rpm, at 28 °C until bacteria grow to an OD_{600} of 0.5. Add 1 ml of bacterial suspension to 100 ml LB medium with the same selective antibiotics in a 250 ml flask and shake on an orbital shaker at 250 rpm, at 28 °C for 4–5 h.
5. Centrifuge at 2500 × *g* for 10 min to collect the bacteria at room temperature. Discard supernatant and resuspend the bacteria in 100 ml of AAM-AS medium.
6. Collect healthy growing light-yellow fragile embryogenic calli into 200 ml sterile flask, and pour AAM-AS medium from **step 5** into the sterile flask, and then add some more AAM-AS medium to immerse the calli for 20–30 min, while shaking occasionally.
7. Dry the excess bacterial suspension pad by drying them on sterile tissue paper, and then place them on a Petri dish with NBD-AS medium covered with two sterile Whatman filter papers #1, soaked with AAM-AS medium. Incubate at 25–28 °C in the dark for 3 days and check for bacterial overgrowth.
8. After 3 days of co-cultivation, wash the calli 5–6 times with sterile water containing 500 mg/L cefotaxime and 500 mg/L carbenicillin and air dry for ~2 h.

9. Transfer the calli evenly to the primary selection medium NBD2 (with 40 mg/L hygromycin and 400 mg/L timentin). Culture the calli for 2 weeks at 25–28 °C in the dark. Culture the calli for 2 more weeks in the same selection medium.
10. Transfer calli to regeneration medium (with 25 mg/L hygromycin and 240 mg/L timentin). Culture under the light at 25–28 °C for around 4 weeks, change the medium every 15 days.
11. Transfer the new shoots into ½ MS (rooting) medium in Magenta box so that the transformed plants can produce roots. Incubate at 25–28 °C, under light.

3.6 Validation of miR166 Silencing by Real-Time PCR Analysis

Transgenic rice expressing STTM166 show very severe phenotype like stunted growth, twisted leaves, and absence of root formation (Fig. 3). Validate the given transgenic lines for knockdown of miR166 and upregulation of target gene expression.

1. Isolate DNA from plants exhibiting severe phenotypes using DNeasy Plant Mini Kit and confirm the transgene integration by STTM specific primers.
2. Isolate RNA from the confirmed plants expressing STTM166 and control vector using Trizol reagent following manufacturer's instructions (*see* **Note 7**).
3. Quantify the RNA samples by measuring the absorbance of the sample at 260 nm using a Nanodrop.

Fig. 3 Transgenic rice expressing STTM166 show very severe phenotype like stunted growth, twisted leaves, and absence of root formation in comparison to the control plants transformed with the empty vector

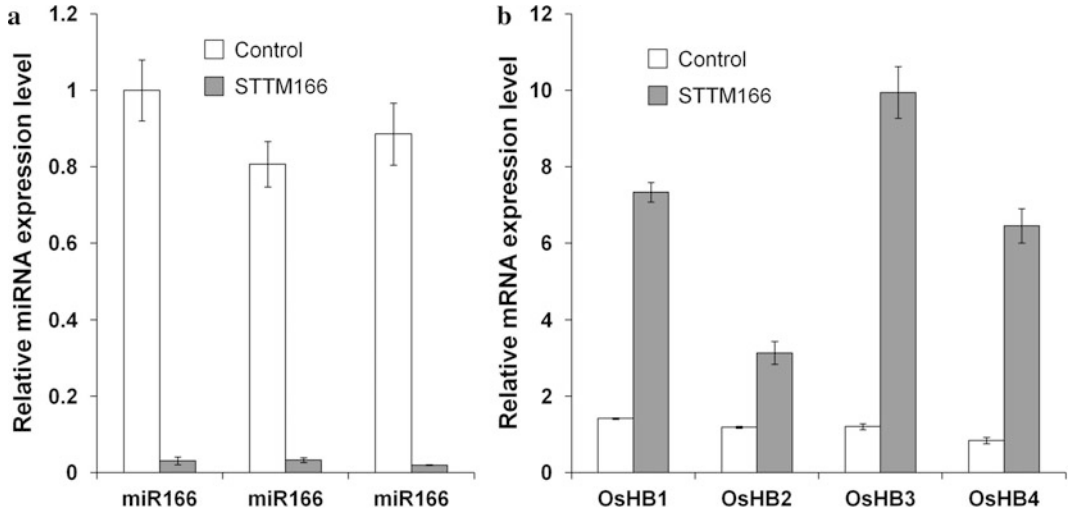


Fig. 4 Real-time PCR analysis of transgenic plants expressing STTM166. (**a**) miR166 expression is suppressed in three independent transgenic lines of rice plants expressing STTM166 construct. (**b**) The target genes of miR166 are upregulated in plants expressing STTM166 construct. The control plants are the wild type plants transformed with the empty vector

4. Check quality of RNA by running in 1% Agarose gel.
5. Take 1 μg of RNA from each independent line and make cDNA using cDNA Reverse Transcription Kit using oligo-dT primers (for target genes) and miR166-stemloop-RT-primer (for miR166). The conditions are as follows: For Stemloop PCR—30 °C for 10 min, 42 °C for 50 min, 85 °C for 5 min; for mRNA RT-PCR—25 °C for 30 min, 42 °C for 30 min, 85 °C for 5 min (*see* **Note 8**).
6. Design real-time PCR primers by selecting a unique region from the target genes and check by using BLAST search engine in NCBI database. The primers should not form a product more than 200 bp.
7. Set up 12 μl of reaction by adding 1 μl of the cDNA, 6 μl of 2× SYBR Green PCR Master mix, 0.5 μl each of 10 μM primers. Set up each reaction in triplicate on a StepOnePlus™ System with the following conditions: heat activation of reverse DNA polymerase at 95 °C for 10 min followed by 40 cycles of 95 °C for 15 s, 60 °C for 1 min.
8. Analyze the data thus obtained by the $2^{\Delta\Delta C_T}$ method [13] (Fig. 4a and b). Take EF1α as an endogenous control and then compare the normalized values to those of control plants.

4 Notes

1. Acetosyringone must always be prepared fresh by dissolving in DMSO (dimethyl sulfoxide) with only one freeze-thaw cycle from −20 °C storage.
2. The tri-nucleotide bulge in the STTM sequence can be other than (CUA) depending upon the mismatches to the complementary bases of the target miRNA.
3. After restriction digestion of pOT2 with *Hin*dIII and *Eco*RI and ligation with the synthetic oligos having the same sites, the final recombinant pOT2 with the STTM will be devoid of both restriction sites.
4. The mentioned PCR conditions were followed for LongAmp® *Taq* DNA polymerase (NEB). Other *Taq* polymerases may have different cycling conditions.
5. In case of silencing miR166, STTM driven by enhanced 35S promoter worked well. Other promoters like Ubiquitin or Actin promoters can also be used, which otherwise work well in the monocots.
6. Before ligating the mixture of PacI-digested pCAMBIA-STTM166 and origin-del-pOT2 vectors, specific dephosphorylation pCAMBIA vector (but never the pOT2-STTM) may be required to increase the colonies bearing recombinant clones on the double antibiotics selection plates. In that case, PacI digestion of the two vectors will be done separately.
7. TRIzol contains phenol and GITC and is hazardous to humans. Always wear a laboratory coat, gloves, and eye protection while handling this solution.
8. The mentioned RT-PCR conditions were used while using High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems). Other kits may have a slight difference in the conditions.

Acknowledgements

This study was partially supported by funds from Henan Agricultural University (HAU) and NSFC (31571679), China. G.T. is supported by the National Science Foundation (NSF; grants IOS-1048216 and IOS-1340001). S.T. is supported by a postdoctoral scholarship from HAU. We are thankful to He Yi and Muhammad Uzair, Shanghai Jiao Tong University, China, for help in rice transformation and Joan Leonard, Ohio State University, Columbus, USA, for help in growing rice in greenhouse.

References

1. Franco-Zorrilla JM, Valli A, Todesco M, Mateos I, Puga MI, Rubio-Somoza I, Leyva A, Weigel D, Garcia JA, Paz-Ares J (2007) Target mimicry provides a new mechanism for regulation of microRNA activity. Nat Genet 39 (8):1033–1037
2. Reichel M, Li Y, Li J, Millar AA (2015) Inhibiting plant microRNA activity: molecular SPONGEs, target MIMICs and STTMs all display variable efficacies against target microRNAs. Plant Biotechnol J 13 (7):915–926
3. Yan J, Gu Y, Jia X, Kang W, Pan S, Tang X, Chen X, Tang G (2012) Effective small RNA destruction by the expression of a short tandem target mimic in Arabidopsis. Plant Cell 24 (2):415–427
4. Sha A, Zhao J, Yin K, Tang Y, Wang Y, Wei X, Hong Y, Liu Y (2014) Virus-based microRNA silencing in plants. Plant Physiol 164(1):36–47
5. Wong J, Gao L, Yang Y, Zhai J, Arikit S, Yu Y, Duan S, Chan V, Xiong Q, Yan J, Li S, Liu R, Wang Y, Tang G, Meyers BC, Chen X, Ma W (2014) Roles of small RNAs in soybean defense against *Phytophthora sojae* infection. Plant J 79 (6):928–940
6. Cao D, Wang J, Ju Z, Liu Q, Li S, Tian H, Fu D, Zhu H, Luo Y, Zhu B (2016) Regulations on growth and development in tomato cotyledon, flower and fruit via destruction of miR396 with short tandem target mimic. Plant Sci 247:1–12
7. Gu Z, Huang C, Li F, Zhou X (2014) A versatile system for functional analysis of genes and microRNAs in cotton. Plant Biotechnol J 12 (5):638–649
8. Jiao J, Wang Y, Selvaraj JN, Xing F, Liu Y (2015) Barley stripe mosaic virus (BSMV) induced microRNA silencing in common wheat (*Triticum aestivum* L.) PLoS One 10 (5):e0126621
9. Teotia S, Singh D, Tang X, Tang G (2016) Essential RNA-based technologies and their applications in plant functional genomics. Trends Biotechnol 34(2):106–123
10. Tang G, Yan J, Gu Y, Qiao M, Fan R, Mao Y, Tang X (2012) Construction of short tandem target mimic (STTM) to block the functions of plant and animal microRNAs. Methods 58 (2):118–125
11. Weigel D, Glazebrook J (2006) Transformation of Agrobacterium using the freeze-thaw method. CSH Protoc 2006(7):pdb.prot4666
12. Nishimura A, Aichi I, Matsuoka M (2006) A protocol for Agrobacterium-mediated transformation in rice. Nat Protoc 1(6):2796–2802
13. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2-\Delta\Delta CT$ method. Methods 25(4):402–408

Chapter 26

RNAi-Mediated Knockdown of Protein Expression

Volker Baumann, Cornelia Lorenzer, Michael Thell, Anna-Maria Winkler, and Johannes Winkler

Abstract

RNA interference is an essential method for studying genomic functions of single genes by loss-of-function experiments. Short interfering siRNAs are efficiently transfected into cultured cells to enable RISC-mediated mRNA cleavage and inhibition of translation in a sequence-specific manner. RNAi enables knockdown of single genes and screening for specific cellular processes or outcomes. In this chapter, we describe a detailed universal protocol for lipoplex-mediated siRNA transfection for cell cultures and cell lysis for subsequent RNA or protein analysis. The experimental procedure is described for verification of knockdown and includes cell lysis for mRNA or protein quantification. Important aspects for specific gene silencing and potential pitfalls are discussed.

Key words siRNA, Gene silencing, Oligonucleotides, Transfection, Lipoplexes, Cell lysis, Off-target effects, Toxicity

1 Introduction

Since the Nobel Prize winning discovery of RNA interference (RNAi) in nematodes [1] and the development of short interfering RNAs (siRNAs) for gene silencing in mammalian cells [2] have been described, RNAi has become a widespread scientific tool for studying gene functions. Incorporation into the intracellular effector RNA-induced silencing complex (RISC) [3] and sequence-specific guiding by the antisense strand of an siRNA duplex to the corresponding mRNA triggers cleavage and degradation of the mRNA, thereby preventing translation into the protein. This powerful technology has found broad use as a scientific tool through loss-of-function experiments. Specific gene knock down can give insight into functional roles of the respective target, and RNAi screening can identify novel genetic players for cellular processes or scientific questions [4]. Intense efforts to translate the RNAi technology into therapeutics have been complicated by poor pharmacokinetic parameters and often insufficient biodistribution and

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_26,

cellular uptake in vivo [5]. Several siRNA oligonucleotides are currently being tested in clinical trials, particularly for liver targets, as this is the organ with the highest accumulation [6, 7]. While either chemical modifications [8] or liposomal or nanoparticle formulations [9–12] are all but required for successful in vivo applications, unmodified siRNA oligonucleotides are highly effective in cell culture after intracellular delivery using transfection reagents. Synthetic siRNA oligonucleotides are most commonly transfected using lipoplexes, cationic peptides or polymers, or with electroporation [13–16]. Transfection procedures are relatively straightforward, but results are dependent on the particular cell line or cell type, as well as on the siRNA target [17, 18]. Sufficient intracellular concentrations of siRNA molecules are essential for robust and efficient gene silencing. For strong and specific target knockdown, transfection should be optimized for the particular cell line and target, and potential toxic effects of the transfection agent need to be monitored. In addition, the possibility of off-target silencing, caused by incomplete binding to similar sequences, needs to be taken into account. Complementarity of nucleotides 2–8 of the siRNA sense strand to other genetic sequences can trigger miRNA-like effects [19, 20]. Thus, the inclusion of proper controls and verification of successful silencing of the target gene are strongly recommended for each new siRNA. This protocol describes the most commonly employed technique, which makes use of lipoplexes, and procedures for preparing cell lysates for subsequent analysis of gene expression at the mRNA and protein level.

2 Materials

2.1 Preparation of siRNA Stock Solution

1. siRNA (single strands or duplex).
2. Nuclease-free water.
3. Phosphate buffered saline solution (PBS), sterile.

2.2 Cell Seeding and Transfection

1. Phosphate buffered saline solution (PBS), sterile.
2. Trypsin solution (0.25% in PBS, for cell culture), or equivalent cell dissociation reagent.
3. Tissue culture plate, 24-well, transparent (*see* **Note 1**).
4. Brightfield microscope.
5. Cell culture incubator (37 °C, 5% CO_2).
6. Laminar flow cabinet.
7. Cell line or primary cells to be transfected, ca. 3×10^6–1×10^7 cells per 24-well plate (*see* **Note 2**).

8. Complete cell culture medium. Use standard cell culture medium, including serum, but without antibiotics, for example, Dulbecco's modified Eagle Medium (DMEM), 10% Fetal Bovine Serum (FBS), 2 mM L-Glutamine (*see* **Note 3**).
9. Lipofectamine RNAiMAX (Life Technologies), or a similar transfection reagent (*see* **Note 4**).
10. siRNA oligonucleotides, duplex, stock solution (1 μM) in PBS (*see* **Note 5**).
11. Opti-MEM (Life Technologies), or other reduced cell culture medium.

2.3 Cell Lysis for RNA Extraction and Analysis

1. Guanidinium acid phenol reagent (Trizol, TRI reagent or similar).
2. Chloroform or 1-bromo-3-chloropropane, for molecular biology.

2.4 Cell Lysis for Protein Analysis

1. Phosphate buffered saline solution (PBS).
2. Lysis buffer for western blotting (RIPA): 50 mM Tris, pH 7–8, 150 mM NaCl, 0.1% SDS, 0.5% sodium deoxycholate, 1% Triton X 100 or NP-40, and 1% protease inhibitor cocktail; or lysis buffer suited for 2D gel electrophoresis: 8 M urea, 2 M thiourea, 0.5% Triton X-100, 2% CHAPS, 5 mM EDTA, 32 mM DTT, and 1% protease inhibitor cocktail.

3 Methods

3.1 Cell Seeding

1. Use cells with around 80% confluency for seeding (*see* **Note 6**). For adherent cells (*see* **Note 7**), in a laminar flow cabinet, wash with PBS and add trypsin solution (1 ml per T-75 flask) for detachment. Incubate for 3–5 min at 37 °C.
2. Check for complete detachment under the microscope. Resuspend cells in pre-warmed complete medium (5–15 ml per T-75 flask), and count cell number in an automated cell counter or using a Neubauer or Thoma chamber (*see* **Note 8**).
3. Dilute cell suspension in complete cell culture medium to 2.5×10^5–1×10^6 per ml (*see* **Note 9**).
4. Dispense 500 μl to each well of a 24-well tissue culture plate. Thus, 1.25×10^5–5×10^5 cells are in each well (*see* **Note 10**).
5. Incubate cells at 37 °C, 5% CO_2 for 18–24 h.

3.2 siRNA Sequence Selection and Preparation of Stock Solution

1. For designing a target sequence specific siRNA duplex, free and commercial software tools are available (*see* **Note 11**). For standard cell culture experiments, unmodified siRNAs (without 2′-modifications) are usually well suited. Candidate

sequences should be checked for specificity and potential off-target effects by BLAST comparison against the transcripts of the used species. If possible, at least three mismatches against mRNAs other than the intended targets should be present.

2. Commercial siRNAs are usually supplied in lyophilized form, or in a stock solution of 1 mM or 100 μM. Dissolve lyophilized siRNA to a 100 μM stock solution in a sterile tube in nuclease-free water or PBS for long-term storage (−70 °C). If separate RNA oligonucleotide strands are provided, mix in an equimolar amount. Heat for 5 min to 80 °C, then cool down to room temperature to ensure complete duplex formation.
3. Dilute stock solution to a 1 μM working solution in Opti-MEM or PBS in a sterile tube (*see* **Note 12**).

3.3 Transfection

1. Pre-warm complete cell culture medium to 37 °C and Opti-MEM to room temperature. Keep Lipofectamine at 4 °C until needed.
2. In a laminar flow cabinet, remove cell culture medium from cells and add 400 μl fresh pre-warmed medium to each well.
3. Dilute the appropriate amount of siRNA in Opti-MEM. For each well, use 0.25–25 pmol siRNA (0.5–50 nM final concentration for transfection) in 50 μl Opti-MEM. So, for triplicate transfection in a 1 nM concentration, dilute 1.75 μl of a 1 μM stock solution in 175 μl Opti-MEM (*see* **Note 13**).
4. Prepare additional siRNA dilutions in the same way for concentration-dependent experiments (*see* **Note 14**) or for additional gene targets. Always include a nonfunctional siRNA (scrambled control or other untargeted sequence) as a control, and at least three replicates of untreated cells (*see* **Note 15**).
5. Dilute the appropriate volume of Lipofectamine RNAiMAX in Opti-MEM. Per pmol of siRNA, 0.3 μl Lipofectamine is diluted in 50 μl Opti-MEM. For a triplicate transfection in a 1 nM concentration, dilute 0.525 μl Lipofectamine in 175 μl Opti-MEM. Mix by thorough shaking or vortexing and spin down.
6. Combine dilutions of siRNA and Lipofectamine by adding siRNA solution into Lipofectamine dilution. Mix by vortexing or thorough shaking and spin down.
7. Incubate for 5 min at room temperature (*see* **Note 16**).
8. Add 100 μl Lipofectamine-siRNA mixtures to the selected wells of the 24-well plate. Add 100 μl Opti-MEM to untreated controls.
9. Incubate cells at 37 °C, 5% CO_2 for 24–72 h (*see* **Note 17**).
10. Optional: Remove cell culture medium after 2–24 h and replace with fresh medium (*see* **Note 18**).

3.4 Cell Lysis for Subsequent RNA Isolation and Analysis

1. Monitor cell growth visually under a microscope (*see* **Note 19**).
2. Remove cell culture supernatant. Sterile conditions are not necessary from this point on.
3. Optional: Wash each well twice with 500 µl PBS.
4. Add 500 µl Trizol to each well (*see* **Note 20**). Work under a hood from this step on (phenol is toxic).
5. Homogenize solution by pipetting a few times, until solution is less viscous.
6. Pipet each sample into a labelled 1.5 ml tube.
7. Add 100 µl chloroform to each tube.
8. Vortex for 3–5 s.
9. Incubate at room temperature for 5 min.
10. Centrifuge at 11,000 × *g*, 15 min, 4 °C.
11. Carefully remove aqueous upper phase (RNA) without aspirating interphase, which contains DNA, and transfer into a fresh labelled tube.
12. Precipitate RNA by adding 300 µl isopropanol (equal volume of aqueous phase) or continue RNA extraction with spin columns according to manufacturer's instructions.
13. For precipitation, cool tube to −20 °C for 30 min and centrifuge for 15 min (r.t., 11,000 × *g*).
14. Wash pellet with 500 µl ice-cold 75% ethanol, centrifuge at 11,000 × *g* for 5 min and discard supernatant. Repeat washing and centrifugation.
15. Air-dry pellet and dissolve RNA in 10–50 µl Tris–HCl (10 mM, pH 7), PBS, or nuclease-free water (*see* **Note 21**). Measure concentration by absorbance determination at 260 nm (*see* **Note 22**).

3.5 Cell Lysis for Protein Analysis

1. Monitor cell growth visually under a microscope (*see* **Note 19**).
2. Put plate on ice.
3. Remove cell culture supernatant. Sterile conditions are not necessary from this point on.
4. Wash each well with 500 µl ice-cold PBS.
5. Add 20–50 µl lysis buffer (RIPA or urea-buffer) directly on top of cell layer (*see* **Note 23**).
6. Incubate on ice for 5 min.
7. Completely detach cells with a cell scraper.
8. Transfer each cell lysate to a labelled tube.
9. Optional: One to three freeze-thaw cycles by immersing in liquid nitrogen and thawing to improve cell lysis.

10. Centrifuge for 3 min at 11,000 × *g* (4 °C) to collect cell debris. Transfer supernatant into fresh tube.
11. Store at 4 °C for short-term or below −20 °C for long-term storage.

4 Notes

1. Cell numbers can be scaled up or down depending on the downstream analyses. Phenotypical characterization, reporter gene readouts, or viability screening can be performed in 96-well plates. For gene expression analysis by qPCR or western blotting, 24-well plates or 12-well plates are recommended. For other applications that require higher cell numbers or cell lysates, 6-well plates or Petri dishes can be used. All amounts and volumes of siRNA, Lipofectamine, and cell culture media can be scaled up linearly according to the cell number.
2. Most cell lines and many primary cells can be successfully transfected with siRNA with this protocol. For some cells that are difficult to transfect, optimization may be required by increasing siRNA concentration, increasing Lipofectamine volume relative to siRNA concentration, or by using serum-free conditions for transfection.
3. In general, it is recommended to use the standard cell growth medium including serum supplementation. Antibiotics interfere with lipoplex formation and thus should not be present during transfection. If necessary, antibiotics can be added after successful transfection.
4. There are many cationic lipid transfection agents available commercially. For most applications, many of those will give a similar transfection efficiency and toxicity.
5. We advise to either design the sequence yourself (*see* **Note 11**) or order siRNAs against the desired target with disclosed nucleotide sequences. Several vendors sell target validated siRNAs, often a mixture of several sequences, but do not disclose the oligonucleotide sequence. This complicates the evaluation of the results, because potential off-targets cannot be readily identified and may cause false positive results. The use of a mixture of three or more siRNA sequences decreases the likelihood of insufficient knockdown, but increases the probability of influencing the expression of an off-target gene (false positive outcome).
6. Use healthy growing cells only and regularly check for mycoplasma contamination. Slowly growing cells and cells that were overgrown in the tissue culture flask prior to seeding result in lower transfection efficiencies. Contamination may interfere

with transfection, cause excess toxicity, and affect the functional consequences of siRNA gene silencing.

7. To adapt this protocol for cells growing in suspension, add centrifugation steps for each medium exchange and before lysis.
8. Ensure optimal cell concentration for the counting method. For manual counting, the cell suspension should be below 1×10^6 per ml. Inaccurate cell counting is a cause for high inter-experimental variation and poor reproducibility.
9. Cell number may need to be adapted for the respective cell line or cell type used. In general, choose the cell number so that a cell confluency of around 90% results at the end of the incubation period, before cell lysis. The indicated range is suitable for most cell lines. For quickly growing cells such as HeLa, seed no more than 1.5×10^5 cells per well. Slowly growing cell lines or primary cells need to be seeded at a higher density (5×10^5 cells per well or higher). If cell number is too low, higher lipofectamine toxicity and poor cell growth may result, finally yielding insufficient cell mass for protein or mRNA analyses. A too high cell number hampers transfection and cell overgrowth may compromise the functional outcomes.
10. In some cases, a reverse transfection protocol may be favorable. To this end, first prepare the transfection mix according to the protocol (3.3) and dispense it to the wells (100 µl each), then add the cell suspension (400 µl). Optimal cell numbers and siRNA concentrations may differ from transfection of an adherent cell layer.
11. Freely available algorithms are siDirect2 (http://sidirect2.rnai.jp/) and the Whitehead tool (http://sirna.wi.mit.edu/home.php) and are mainly based on thermodynamic stability and specific sequence design rules. Commercial vendors offering siRNA design tools include GE Life Sciences/Dharmacon (http://dharmacon.gelifesciences.com/design-center/), InVivoGen (http://www.invivogen.com/sirnawizard/design.php), and GenScript (http://www.genscript.com/siRNA_target_finder.html). Registration may be required for commercial sites. If not included in the design tool itself, it is strongly recommended to check the putative siRNA sequence for unintended binding to non-targets using nucleotide BLAST (http://blast.ncbi.nlm.nih.gov/). Search the refseq_rna database for the target organism with megablast parameters to identify potential off-targets. The selected siRNA sequence should have at least two to three mismatches against other mRNAs, preferably within nucleotides 2–9 of the siRNA.

    These design tools generally yield around 70–80% of active siRNAs (defined by knockdown of the mRNA to 30% or less of

the original level). However, for highly efficient knockdown of a particular target gene, the evaluation of two or three potential siRNA sequences may be necessary.

12. This working solution concentration is suited for 0.5–10 nM concentrations in 24-well plates. For smaller or larger areas or for higher concentrations, other working solutions may be required.
13. These instructions are for triplicate experiments with a 50 µl surplus volume. For a higher number of replicates, scale up accordingly. For concentration-dependent experiments, a lipoplex mixture of the highest concentration can be produced and diluted after completed complexation of siRNA and Lipofectamine.
14. siRNA concentrations between 1 and 5 nM give efficient target knockdowns for most targets and cell lines. In some cases, higher concentrations up to 50 nM may be necessary. However, keep in mind that high concentrations increase the risk of off-target effects caused by binding to targets with non-perfect sequence complementarity. In addition, lipofectamine toxicity is concentration- and cell line dependent. The lowest siRNA concentration that gives efficient gene knockdown should be used. For new targets and cell lines, it is strongly recommended to verify successful gene silencing by determining mRNA or protein levels and monitoring the absence of unspecific toxicity related to different lipofectamine concentrations (*see* **Note 18**).
15. These controls are essential for assessing unspecific effects caused by the transfection reagent or the eventual chemical modification of the oligonucleotides [21, 22].
16. A minimal incubation time of 5 min is required to complete lipoplex formation. Incubation can be increased up to 1 h without significant loss of transfection efficiency, however longer incubation times should be avoided and lipoplexes should be prepared each time directly before use. If a higher number of siRNAs are used, we recommend preparing all lipoplexes and starting the 5 min incubation when the last sample is finished.
17. Incubation time depends mainly on the downstream readout or experiments. mRNA knockdown generally starts already several hours after start of transfection and is active at least for 72 h, often longer if chemically stabilized siRNA is used. Functional loss of protein activity is also dependent on protein half-life, as only the translation is inhibited.
18. The majority of lipoplexes are taken up into cells during the first 1–2 h after addition to the cell culture medium. If needed for cell health or if antibiotics are required for cell maintenance, the cell culture medium can be replaced after 2–4 h or at a later time point without loss of transfection efficiency.

19. We advise to regularly check for cell viability and cell health visually during the incubation period. Reduced cell numbers or a larger number of non-viable cells may be caused by a too-high lipofectamine concentration. Depending on the siRNA target, cell death may also be triggered by gene silencing. Comparison to control siRNA and untreated samples gives an indication of toxicity and whether a reduction of lipoplex concentration is required. For detailed evaluation of specific or unspecific toxicity, cell counting or cell viability assessment is recommended.
20. As an alternative, commercial kits for RNA isolation can be used. Lyse cells in lysis buffer and progress according to the manual.
21. It is crucial to completely remove the ethanol, as it interferes with downstream analyses and inhibits reverse transcriptases. However, the pellet should both not be completely dry, because it complicates dissolving.
22. Besides absorbance readings at 260 nm, maximum wavelength and the absorbance ratio between A260/A280 should be observed. A260/A280 ratio is usually around 2.0, and significantly lower values are indicative of protein contamination. A peak shift towards 270 nm is caused by phenol contamination, which can be removed by repeated butanol extraction [23].
23. Optimal lysis buffer volume is dependent on cell number and the protein content of the cell line. Typically, 25 µl gives protein concentrations between 2 and 5 µg/µl.

References

1. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391 (6669):806–811
2. Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, Tuschl T (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature 411 (6836):494–498
3. Pratt AJ, MacRae IJ (2009) The RNA-induced silencing complex: a versatile gene-silencing machine. J Biol Chem 284(27):17897–17901. doi:10.1074/jbc.R900012200
4. Sharma S, Rao A (2009) RNAi screening: tips and techniques. Nat Immunol 10(8):799–804
5. Dirin M, Winkler J (2013) Influence of diverse chemical modifications on the ADME characteristics and toxicology of antisense oligonucleotides. Expert Opin Biol Ther 13(6):875–888. doi:10.1517/14712598.2013.774366
6. Nair JK, Willoughby JLS, Chan A, Charisse K, Alam MR, Wang Q, Hoekstra M, Kandasamy P, Kel'in AV, Milstein S, Taneja N, O'Shea J, Shaikh S, Zhang L, van der Sluis RJ, Jung ME, Akinc A, Hutabarat R, Kuchimanchi S, Fitzgerald K, Zimmermann T, van Berkel TJC, Maier MA, Rajeev KG, Manoharan M (2014) Multivalent N-acetylgalactosamine-conjugated siRNA localizes in hepatocytes and elicits robust RNAi-mediated gene silencing. J Am Chem Soc 136(49):16958–16961. doi:10.1021/ja505986a
7. Lorenzer C, Dirin M, Winkler A-M, Baumann V, Winkler J (2015) Going beyond the liver: progress and challenges of targeted delivery of siRNA therapeutics. J Control Release 203:1–15. doi:10.1016/j.jconrel.2015.02.003
8. Dirin M, Urban E, Lachmann B, Noe CR, Winkler J (2015) Concise postsynthetic preparation of oligonucleotide-oligopeptide conjugates through facile disulfide bond formation.

Future Med Chem 7(13):1657–1673. doi:10.4155/fmc.15.109

9. Gallas A, Alexander C, Davies MC, Puri S, Allen S (2013) Chemistry and formulations for siRNA therapeutics. Chem Soc Rev 42 (20):7983–7997. doi:10.1039/C3CS35520A
10. Winkler J (2013) Oligonucleotide conjugates for therapeutic applications. Ther Deliv 4 (7):791–809. doi:10.4155/tde.13.47
11. Winkler J (2015) Therapeutic oligonucleotides with polyethylene glycol modifications. Future Med Chem 7(13):1721–1731. doi:10.4155/fmc.15.94
12. Winkler J (2011) Nanomedicines based on recombinant fusion proteins for targeting therapeutic siRNA oligonucleotides. Ther Deliv 2 (7):891–905. doi:10.4155/tde.11.56
13. Winkler J, Gilbert M, Kocourkova A, Stessl M, Noe C (2008) 2′-O-lysylaminohexyl oligonucleotides: modifications for antisense and siRNA. ChemMedChem 3(1):102–110. doi:10.1002/cmdc.200700169
14. Winkler J, Saadat K, Diaz-Gavilan M, Urban E, Noe C (2009) Oligonucleotide-polyamine conjugates: influence of length and position of 2′-attached polyamines on duplex stability and antisense effect. Eur J Med Chem 44(2):670–677. doi:10.1016/j.ejmech.2008.05.012
15. Dirin M, Urban E, Noe CR, Winkler J (2016) Fragment-based solid-phase assembly of oligonucleotide conjugates with peptide and polyethylene glycol ligands. Eur J Med Chem 121: 132–142. doi:10.1016/j.ejmech.2016.05.001
16. Höbel S, Aigner A (2013) Polyethylenimines for siRNA and miRNA delivery in vivo. Wiley Interdiscip Rev Nanomed Nanobiotechnol 5 (5):484–501. doi:10.1002/wnan.1228
17. Gaziova Z, Baumann V, Winkler A-M, Winkler J (2014) Chemically defined polyethylene glycol siRNA conjugates with enhanced gene silencing effect. Bioorg Med Chem 22(7):2320–2326. doi:10.1016/j.bmc.2014.02.004
18. Shokrzadeh N, Winkler A-M, Dirin M, Winkler J (2014) Oligonucleotides conjugated with short chemically defined polyethylene glycol chains are efficient antisense agents. Bioorg Med Chem Lett 24(24):5758–5761. doi:10.1016/j.bmcl.2014.10.045
19. Doench J, Petersen C, Sharp P (2003) siRNAs can function as miRNAs. Genes Dev:438–442. doi:10.1101/gad.1064703.et
20. Baumann V, Winkler J (2014) miRNA-based therapies: strategies and delivery platforms for oligonucleotide and non-oligonucleotide agents. Future Med Chem 6(17):1967–1984. doi:10.4155/fmc.14.116
21. Stessl M, Marchetti-Deschmann M, Winkler J, Lachmann B, Allmaier G, Noe C (2009) A proteomic study reveals unspecific apoptosis induction and reduction of glycolytic enzymes by the phosphorothioate antisense oligonucleotide oblimersen in human melanoma cells. J Proteome 72(6):1019–1030. doi:10.1016/j.jprot.2009.06.001
22. Winkler J, Stessl M, Amartey J, Noe C (2010) Off-target effects related to the phosphorothioate modification of nucleic acids. ChemMedChem 5(8):1344–1352. doi:10.1002/cmdc.201000156
23. Krebs S, Fischaleck M, Blum H (2009) A simple and loss-free method to remove TRIzol contaminations from minute RNA samples. Anal Biochem 387(1):136–138. doi:10.1016/j.ab.2008.12.020

Chapter 27

Engineered Zinc Finger DNA-Binding Domains: Synthesis, Assessment of DNA-Binding Affinity, and Direct Protein Delivery to Mammalian Cells

Mir A. Hossain, Isaac J. Knudson, Shaleen Thakur, Yong Shen, Jared R. Stees, Joeva J. Barrow, and Jörg Bungert

Abstract

Zinc finger proteins are the most common among families of DNA-binding transcription factors. Designer transcription factors generated by the fusion of engineered zinc finger DNA-binding domains (ZF-DBDs) to effector domains have been valuable tools for the modulation of gene expression and for targeted genome editing. However, ZF-DBDs without effector domains have also been shown to effectively modulate gene expression by competing with sequence-specific DNA-binding transcription factors. Here, we describe the methodology and provide a detailed workflow for the cloning, expression, purification, and direct cell delivery of engineered ZF-DBDs. Using this protocol, ZF-DBDs can be generated with high efficiency in less than 2 weeks. We also describe a nonradioactive method for measuring DNA binding affinity of the purified ZF-DBD proteins as well as a method for direct delivery of the purified ZF-DBDs to mammalian cells.

Key words Engineered zinc finger DNA-binding domain, Modular assembly, Recombinant protein purification, Electrophoretic mobility shift assay, Direct protein delivery

1 Introduction

The zinc finger is the most common structural feature in mammalian DNA-binding proteins. Recent advances in understanding the mode of zinc finger-DNA interactions led to the ability to engineer synthetic zinc finger proteins to target specific genomic DNA sequences [1]. Zinc fingers are small protein domains consisting of an α helix, an antiparallel β sheet, and a structural zinc ion. Each zinc finger binds to 3 bp in the major groove of the DNA via interactions between specific amino acid residues in the α helix and the DNA bases. Multiple zinc fingers are linked in an array with flexible linkers to generate the DNA-binding domain of natural or synthetic zinc finger DNA-binding proteins [2]. For synthetic

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_27,

proteins an array of six zinc fingers is usually designed which binds an 18 bp target DNA sequence. A DNA sequence of 18 bp or longer statistically specifies a unique sequence in the mammalian genome [3]. Engineered zinc finger DNA-binding domains (ZF-DBDs) have been fused with various effector domains for genome editing, epigenome remodeling, and for modulation of gene expression [3, 4]. ZF-DBDs without effector domains are also becoming a powerful tool to modulate gene expression [5–7]. Recent studies have shown that engineered ZF proteins can be directly delivered into mammalian cells [8, 9]. The small molecular size and the intrinsic cell penetrating abilities due to high positive charge at the surface render ZF proteins potential candidates for therapeutic interventions. Various methods have been developed to design and select for ZF-DBDs including modular assembly, Oligomerized Pool Engineering (OPEN), and bacterial two-hybrid selection [2]. All these methods have been described in detail in previous publications [10–13]. Here, we describe efficient protocols for the generation, purification, affinity measurement, and direct protein delivery into mammalian cells of engineered ZF-DBDs. In Subheading 3.1, we describe a step-by-step protocol for the generation of a six zinc finger DNA-binding domain (6 ZF-DBD) using a PCR-based modular assembly method modified from a previous publication [14]. Following this method ZF-DBDs can successfully be synthesized within 2 weeks using standard molecular biology techniques. In Subheading 3.2, we provide a protocol for recombinant protein expression and purification. In Subheading 3.3, we outline a protocol for quantitative measurement of the in vitro binding affinity of the generated ZF-DBD proteins. In Subheading 3.4, we describe a protocol for direct delivery of the purified ZF-DBD proteins into mammalian cells.

2 Materials

2.1 Equipment

1. Incubators (37 °C) for growth of bacteria and mammalian cells.
2. Equipment and reagents for mammalian cell culture (biosafety hood, culture flasks, multidishes, hemocytometer, trypan blue).
3. Shaker (37 °C) for bacterial culture.
4. Shaker (18 °C) for recombinant protein expression.
5. Heat blocks (95 °C).
6. Spectrophotometer for measuring optical density of bacterial culture and DNA concentrations.
7. Thermal cycler.
8. Sonicator for lysis of bacterial cells.

9. Temperature-controlled centrifugation system for large volume (10–50 ml) samples.
10. Temperature-controlled tabletop centrifugation system for small volume (1.5–2 ml) samples.
11. Equipment and reagents for agarose and polyacrylamide gel electrophoresis.
12. Western blotting equipment and detection reagents.
13. 6% TBE gels (15-well) for gel shift assays.
14. PCR purification kit.
15. Plasmid DNA extraction kit.
16. Gel extraction kit.
17. NE-PER nuclear and cytoplasmic extraction kit (Pierce).
18. Bio-Rad protein assay.
19. Nickel resin and column for protein purification.
20. Vivaspin 6 concentrators (GE healthcare).
21. Typhoon 9410 imager.
22. ImageJ software.
23. GraphPad Prism software.

2.2 Media, Buffers, and Reagents

1. Standard Lysogeny Broth (LB) medium.
2. LB agar plates with 100 μg/ml ampicillin.
3. Growth medium for culturing K562 cells: RPMI 1640 medium, 10% fetal bovine serum, 1% penicillin/streptomycin.
4. IPTG (isopropyl-β-D-thio-galactopyranoside), 1 M stock prepared in ddH_2O, filter sterilized, and stored at −20 °C.
5. $ZnCl_2$, 100 mM stock prepared in ddH_2O, filter sterilized, and stored at −20 °C.
6. Ampicillin, 100 mg/ml stock prepared in ddH_2O, filter sterilized, and stored at −20 °C.
7. Complete protease inhibitor cocktail, 50× stock prepared in ddH_2O, and stored at −20 °C.
8. Bovine serum albumin stock solution (10 mg/ml).
9. TE buffer (pH 6.8 and pH 8.0).
10. dNTP mix.
11. 5× Phusion HF buffer.
12. NEB 10× CutSmart buffer.
13. DNA modification enzymes (Phusion high fidelity DNA polymerase, restriction enzymes, T4 DNA ligase, calf intestinal phosphatase).
14. Bacteria lysis buffer: 100 mM HEPES, 10 mM imidazole, pH 7.5.

15. Wash buffer 1: 100 mM HEPES, 250 mM imidazole, pH 7.5.
16. Wash buffer 2: 100 mM HEPES, 500 mM imidazole, pH 7.5.
17. Elution buffer: 100 mM HEPES, 1 M imidazole, pH 7.5.
18. ZF-DBD protein storage buffer: 50 mM Tris, pH 7.5, 150 mM NaCl, 100 μM $ZnCl_2$, 1 mM DTT, 10% glycerol.
19. Herring sperm DNA stock solution (10 mg/ml).
20. 5× EMSA binding buffer: 50 mM Tris, pH 7.5, 250 mM KCl, 25 mM $MgCl_2$, 500 μM $ZnCl_2$, 5 mM DTT, 0.25% Triton X-100, and 12.5% glycerol.
21. 0.5× TBE running buffer: 45 mM Tris, pH 8.3, 45 mM boric acid, 1 mM EDTA, and 100 μM $ZnCl_2$.
22. 6× Orange-G loading dye: 10 mM Tris, pH 7.5, 15% Ficoll-400, 0.1% Orange-G.
23. Coomassie stain solution: 50% methanol, 10% acetic acid, 0.25% coomassie blue R-250.
24. Destain solution: 50% methanol, 10% acetic acid.
25. Heparin wash buffer: 1× PBS, 0.5 mg/ml heparin.
26. Primary antibodies (zinc finger antisera, anti-GAPDH, CST5174, Cell Signaling, Beverly, MA, and anti-CTCF, 2899, Cell Signaling, Beverly, MA).
27. Secondary antibody (anti-rabbit IgG HRP).

2.3 Bacterial Strains, Mammalian Cells, Vectors, and Primers

1. *E. coli* strains Stbl2 and BL21(DE3).
2. Mammalian cells (K562 cells).
3. Bacterial expression vector pT7-FLAG-2.
4. Single-stranded 5′ Cy5-labeled oligonucleotide containing the 6 ZF-DBD target site: 5′-GAGAACTTAAGAGA-TAATGGCCTAAAACCACAGAGAGTATAT-3′ (*see* **Note 1**).
5. Single-stranded unlabeled oligonucleotide containing the reverse complement sequence of the Cy5-labeled oligonucleotide.
6. Oligonucleotides for assembly of ZF constant backbones. C1 (ZF1–3): 5′-CGGGGAGAAACCCTATAAGTGTCCGGAGT GTGGCAAGTCGTTCTC-3′; C1(ZF4–6): 5′-TATAAGTGT CCGGAGTGTGGCAAGTCGTTCTC-3′; C2(ZF1–3 and ZF4–6): 5′-GCGTACCCACACGGGCGAAAAGCCGTACA AATGCCCAGAATGCGGTAAATCCTTCAGC-3′; C3 (ZF1–3 and ZF4–6): 5′-TCAACGGACGCATACAGGAG AGAAGCCATACAAATGTCCCGAATGTGGGAAGAGTTT TAG-3′.
7. Oligonucleotides for assembly of ZF variable regions (finger specific). V1, V2, V3, V4, V5, and V6 (*see* **Note 1**).

8. Primers for amplification of ZF1–3 and ZF4–6 assembly products. ZF1–3(His-*HindIII*)F: 5′-CAGGACAAGCTTCAC CACCACCACCACCACCTCGAGCCCGGGGAGAAACCC TATAAG-3′. ZF1–3(*AgeI*)R: 5′-CAGGACACCGGTGTGA GTGCGCTGGTG -3′. ZF4–6(*AgeI*)F: 5′-GACACCGGTG GTGGCGGAGGTGAACGAGAGAAGCCCTATAAGTGTC CGGAGTGTGG-3′. ZF4–6(*BglII*)R: 5′-CAGGACAGATC TTCAGCTGGTTTTTTTGCCGGTGT-GAGTGCGCTGGTG-3′.
9. Sequencing primer for pT7 vector. pT7 primer: 5′-TAATAC GACTCACTATAGGG-3′.

3 Methods

The ZF-DBD used in this study was designed using a free online tool (zincfingertools.org) which utilizes knowledge from ZF-DNA interactions in the context of 3 ZF subsets [10]. The zinc fingers designed in this manner contain canonical amino acid linkers (TGEKP) between the ZFs of the 3 ZF modules (ZF 1–3 and ZF 4–6) and an extended linker (TGGGGGEREKP) between the modules. The amino acid sequences of the zinc finger array were provided by the zinc finger tool and were used to design oligonucleotides to assemble the ZF-DBD coding sequence using a modified version of PCR assembly described previously [5, 14]. We found that the cloning efficiency of 6 ZF-DBD was greatly improved when we followed a two-step cloning strategy (Fig. 1). The ZF-DBD was expressed in *E. coli* and was subsequently purified using nickel columns. Purified ZF-DBD proteins were subjected to DNA-binding affinity measurement in vitro using electrophoretic mobility shift assays (EMSAs) with a fluorescently labeled double-stranded DNA oligonucleotide. Furthermore, we adopted a protocol from Gaj et al. 2012 [15] and directly delivered the purified ZF-DBD protein into K562 cells.

3.1 PCR Assembly and Cloning of ZF-DBD

1. Prepare working dilutions of the oligonucleotides for the ZF-DBD constant backbones (C1, C2, and C3) and variable regions (V1, V2, V3, V4, V5, and V6) to final concentrations of 25 μM each. Dilute the amplification primers to final concentrations of 10 μM each.
2. Prepare an assembly reaction for ZF1–3 containing 1 μl of each of the 25 μM C1(ZF1–3), C2, C3, V1, V2, V3.
3. Similarly, prepare another assembly reaction for ZF4–6 containing 1 μl of each of the 25 μM C1(ZF4–6), C2, C3, V4, V5, V6.
4. Add 1 μl of 10 mM dNTP mix, 10 μl 5× Phusion HF buffer, and 1 U of Phusion high fidelity DNA polymerase to each of

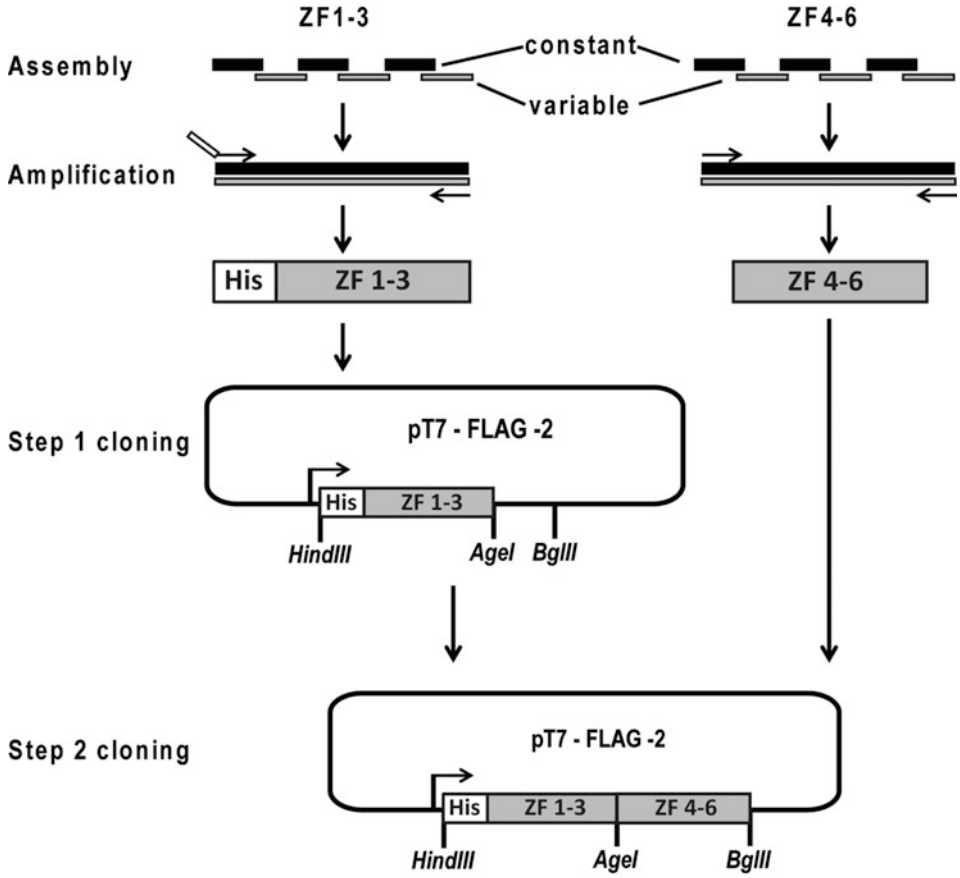


Fig. 1 Strategy for the generation of 6 ZF-DBDs. Oligonucleotides for the ZF constant backbones and variable regions were assembled and were subsequently amplified to generate 3 ZF modules. The ZF modules were cloned into pT7-FLAG-2 vector using a two-step cloning strategy. In step 1, ZF1–3 was inserted into the vector using the indicated restriction enzyme sites. In step 2, ZF4–6 was inserted into the vector containing the correct sequence of ZF1–3 using the indicated restriction enzyme sites

the reactions. Bring the final volume to 50 μl with nuclease-free water.

5. Perform assembly PCR using the following condition: initial denaturation at 95 °C for 2 min, followed by 12 cycles of 30 s at 95 °C, 30 s at 58 °C, and 30 s at 72 °C, followed by a final extension at 72 °C for 5 min.
6. Clean up the assembly PCR products using a PCR purification kit with an elution volume of 30 μl each.
7. Prepare amplification PCR reaction for ZF1–3 by combining 1 μl of the elution from **step 6**, 1 μl of each of the 10 μM forward ZF1–3(His-*HindIII*)F and reverse ZF1–3(*AgeI*)R primers, 1 μl of 10 mM dNTP mix, 10 μl 5× Phusion HF buffer, and 1 U of Phusion high fidelity DNA polymerase. Bring the final volume to 50 μl with nuclease-free water.

8. Similarly, prepare another amplification PCR reaction for ZF4–6 by combining 1 μl of the elution from **step 6**, 1 μl of each of the 10 μM forward ZF4–6(*AgeI*)F and reverse ZF4–6 (*BglII*)R primers, 1 μl of 10 mM dNTP mix, 10 μl 5× Phusion HF buffer, and 1 U of Phusion high fidelity DNA polymerase. Bring the final volume to 50 μl with nuclease-free water.
9. Perform amplification PCR with the following program: initial denaturation at 95 °C for 2 min, followed by 27 cycles of 30 s at 95 °C, 30 s at 58 °C, and 30 s at 72 °C, followed by a final extension at 72 °C for 5 min.
10. Clean up the amplification PCR products using a PCR purification kit with an elution volume of 30 μl each.
11. Digest eluent of ZF1–3 with *HindIII* and *AgeI*, and eluent of ZF4–6 with *AgeI* and *BglII*.
12. Subject digested samples to electrophoresis using 1.8% agarose gels and stain with ethidium bromide. Using UV shadowing, cut bands corresponding to 300 bp with a clean sharp razor blade. Purify the DNA using a gel extraction kit with an elution volume of 30 μl each.
13. Ligate the digested ZF1–3 DNA fragment into the *HindIII* and *AgeI* sites of the pT7-FLAG-2 vector (*see* **Note 2**).
14. Verify the sequence integrity of the cloned product by Sanger sequencing using pT7 primers.
15. Purify plasmid DNA from the clone with the correct sequence of ZF1–3, digest with *AgeI* and *BglII*, and re-purify using a PCR purification kit.
16. Ligate the digested ZF4–6 DNA fragment (from **step 12**) into the digested vector containing the ZF1–3 (from **step 15**) (*see* **Note 3**).
17. Verify the sequence integrity of the cloned product (ZF1–6) by Sanger sequencing using pT7 primers (*see* **Note 4**).

3.2 Expression and Purification of ZF-DBD

1. Transform pT7-FLAG-2 vector containing ZF1–6 sequence into BL21(DE3) competent cells. Streak out on LB agar plates with 100 μg/ml ampicillin and grow at 37 °C overnight.
2. On the next day, pick one colony and start an overnight culture in 3 ml LB medium with 100 μg/ml ampicillin at 37 °C.
3. Dilute the overnight culture 100 times by transferring 2 ml of overnight culture into 200 ml of fresh LB medium with 100 μg/ml ampicillin and grow at 37 °C with 250 rpm until OD_{600} is ~0.5 (typically 2–3 h).
4. Add 200 μl of 1 M IPTG (1 mM final concentration) and 200 μl of 100 mM $ZnCl_2$ (100 μM final concentration), and grow at 18 °C with 250 rpm overnight (*see* **Note 5**).

5. Harvest the bacterial cells by centrifugation at 5000 × *g* for 10 min at 4 °C. Resuspend the pellet in 3 ml of ice-cold bacteria lysis buffer with 1× complete protease inhibitor cocktail.
6. Break the bacterial cells using a sonicator at high power setting for 15 cycles with 30 s on/30 s off at 4 °C (*see* **Note 6**).
7. Centrifuge the lysates at 18,000 × *g* for 15 min at 4 °C and transfer the supernatant into a fresh tube. Keep on ice.
8. Add 2 ml of resuspended nickel resin (50% slurry) to a 10 ml column. Equilibrate the column by washing with 10 ml of lysis buffer. Stop the flow using a stopper (*see* **Note 7**).
9. Gently pipette the protein lysates on top of the resin bed and let the resin settle down.
10. Open the stopper and collect the flow-through. Wash the resin with 10 ml each of ice-cold lysis buffer, wash buffer 1, and wash buffer 2, respectively.
11. Elute proteins with 10 ml of ice-cold elution buffer.
12. Centrifuge the elution fractions at 18,000 × *g* for 10 min at 4 °C. Transfer the supernatant to a fresh Vivaspin 6 concentrator with a 3 kDa molecular weight cutoff. Desalt and concentrate purified proteins as desired following the protocol from the manufacturer (*see* **Note 8**).
13. Store the protein at −80 °C in ZF-DBD protein storage buffer with 1× complete protease inhibitor cocktail.
14. Prepare dilutions of BSA stock to achieve 50, 100, 150, and 200 ng of BSA.
15. Electrophorese the BSA standards alongside with different amounts of the purified ZF-DBD protein stock (1–2 μl) in 4–15% SDS-PAGE.
16. Stain the gel with Coomassie stain and subsequently destain.
17. Take photograph of the gel and measure the band intensities of all the samples (BSA and ZF-DBD) using ImageJ software. Prepare a standard curve using band intensities (arbitrary unit) and protein amount (ng) of BSA. Calculate the concentration of the ZF-DBD protein stock using the standard curve and band intensities of ZF-DBD proteins (Fig. 2) (*see* **Note 9**).

3.3 Determination of ZF-DBD DNA-Binding Affinity Using EMSAs

1. Prepare Cy5-labeled oligonucleotide in TE pH 6.8 buffer, and unlabeled reverse complement oligonucleotide in TE pH 8.0 buffer to final concentrations of 100 μM each (*see* **Note 10**).
2. Combine 5 μl of Cy5-labeled oligonucleotide and 10 μl of reverse complement with 185 μl of TE pH 8.0 buffer (*see* **Note 11**).
3. Heat the mixture at 95 °C for 5 min. Turn off the heat and let the sample cool down to RT slowly in the heat block.

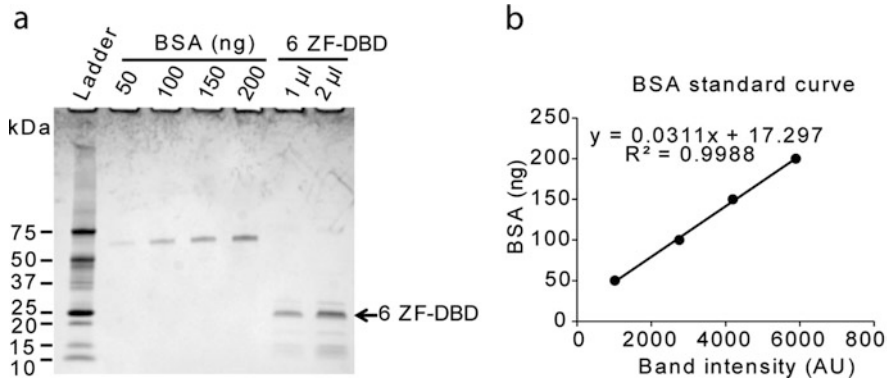


Fig. 2 Measurement of purified 6 ZF-DBD concentration using SDS-PAGE. (**a**) BSA (50–200 ng) and purified 6 ZF-DBD (1–2 µl) were electrophoresed in a 4–15% Tris-glycine SDS polyacrylamide gel. The gel was stained with Coomassie blue. Band intensities for BSA (~66 kDa) and the 6 ZF-DBD (~25 kDa) were measured using ImageJ software. (**b**) The BSA standard curve was obtained by plotting the band intensities (AU, arbitrary unit) against the protein amount (ng) of BSA. The concentration of 6 ZF-DBD (~150 ng/µl) was measured using the equation shown in the graph

4. Measure the concentration of the double-stranded DNA probe using a spectrophotometer.
5. Prepare aliquots of 10 µl and store at −20 °C in a dark container or in Eppendorf tubes wrapped with aluminum foil.
6. Prepare a master mix for 15 EMSA reactions with 1.5 µg of herring sperm DNA, 60 µl 5× EMSA binding buffer, 2.9 pmol of Cy5-labeled probe (for 10 nM final concentration), and ddH_2O in a total volume of 270 µl. Keep on ice.
7. Prepare 11 series of dilutions of ZF-DBD protein stock ranging from 10 nM to 1 µM in ddH_2O with at least 5 µl volume of each dilution. Keep on ice.
8. Prepare 12 reaction tubes with 18 µl master mix in each and label 11 of them according to the dilution series of the ZF-DBD protein. Label one tube as control.
9. Add 2 µl of each ZF-DBD protein dilutions to their corresponding reaction tubes. Add 2 µl ddH_2O to the control tube.
10. Vortex the tubes to mix and spin quickly to bring the volumes to the bottom of the tubes. Incubate in the dark at RT for 30 min.
11. In the meantime, prepare the electrophoresis chamber with the 6% TBE gel and 0.5× TBE running buffer. Run the gel for at least 20 min at RT at constant 50 V (this pre-run equilibrates the gel to the buffer system).
12. Load 3 µl of 6× Orange-G loading dye in the first lane to track the samples during the gel electrophoresis. Load 10 µl samples

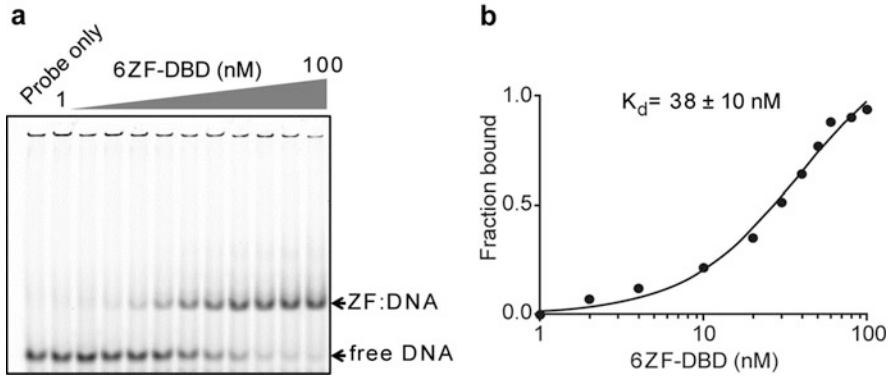


Fig. 3 In vitro DNA-binding affinity measurement of the purified 6 ZF-DBD. (**a**) EMSA showing the signal from the Cy5-labeled DNA probe and that of the protein/DNA complexes. Concentrations of 1–100 nM of the ZF-DBD were used in the experiment. (**b**) Binding curve generated from panel **a** using the fraction of bound DNA plotted against the ZF-DBD concentrations. The fraction of DNA bound was calculated by dividing the signal of the protein/DNA complexes with the combined signals (bound plus free). The dissociation constant (K_d) of the ZF-DBD was determined to be 38 ± 10 nM

from the reaction tubes into each well. Electrophorese at 4 °C at constant 50 V for 2–3 h (*see* **Note 12**).

13. Carefully remove the gel from the cassette and visualize the Cy5-labeled DNA band using a Typhoon 9410 imager (*see* **Note 13**).
14. Measure the band intensities of the bound DNA and free DNA for each lane using ImageJ software.
15. Calculate the fraction bound by dividing the value corresponding to bound DNA with the value corresponding to total DNA (bound plus free).
16. Plot the DNA-fraction bound against the molar concentrations of the ZF-DBD (in this case 1–100 nM). Using Graphpad prism perform a curve fitting for nonlinear regression. This curve fitting provides the dissociation constant (K_d) of the ZF-DBD that defines the amount of protein needed to achieve half maximal binding (Fig. 3) (*see* **Note 14**).

3.4 Direct Protein Delivery of ZF-DBD to Mammalian Cells

1. Culture K562 cells in RPMI medium with 10% FBS and 1% penicillin/streptomycin at a density of 1×10^6 cells/ml with 5% CO_2 at 37 °C. Culture for at least 5 days with a minimum of two passages before initiating the protein delivery experiments (*see* **Note 15**).
2. On the day of the experiment, count cells using a hemocytometer and pipette 2×10^6 cells into an Eppendorf tube.
3. Centrifuge at 200 × *g* for 5 min at RT. Resuspend cell pellet in 1 ml of serum-free medium (base RPMI medium). Divide the suspension into two Eppendorf tubes with 500 μl each (*see* **Note 16**).

4. Add the ZF-DBD protein stock to one of the tubes at a final ZF-DBD concentration of 100 nM. Add ZF-DBD protein storage buffer to the other tube (same volume as with ZF-DBD protein stock) (*see* **Note 17**).
5. Mix thoroughly by pipetting, and transfer the contents of each tube to their corresponding well in a 4-well multidish plate. Incubate at 37 °C for 1 h.
6. Transfer the cells from the multidish plate to fresh Eppendorf tubes. Centrifuge at 200 × *g* for 5 min at RT. Discard the supernatant.
7. Add 500 μl of heparin wash buffer to each tube and gently pipette to resuspend the pellet (*see* **Note 18**).
8. Centrifuge at 200 × *g* for 5 min at RT and discard the supernatant.
9. Repeat **steps 7** and **8** once more. Keep the final cell pellet on ice.
10. Perform nuclear and cytoplasmic protein fractionation using NE-PER kit. Measure the protein concentrations using Bio-Rad protein assay (*see* **Note 19**).
11. Electrophorese 10 μg of each nuclear and cytoplasmic extract in 4–15% SDS-PAGE and perform Western blotting using zinc finger antisera, anti-GAPDH, and anti-CTCF antibodies (Fig. 4) (*see* **Note 20**).

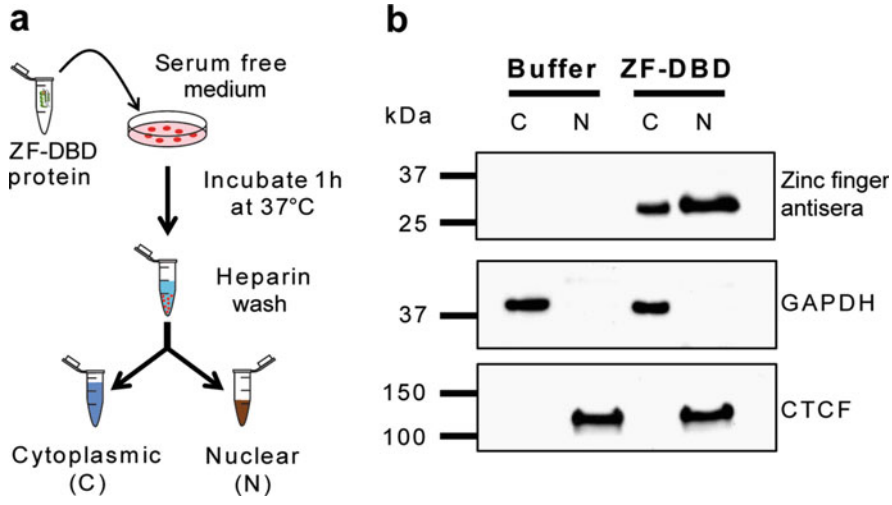


Fig. 4 Direct delivery of the 6 ZF-DBD to K562 cells. (**a**) General outline of the direct protein delivery strategy. (**b**) K562 cells were incubated either with the purified 6 ZF-DBD (100 nM) or with ZF protein storage buffer (Buffer) in serum-free medium for 1 h at 37 °C. Cytoplasmic (C) and nuclear (N) proteins were extracted and analyzed using Western blotting with zinc finger antisera, anti-GAPDH, or anti-CTCF antibodies

4 Notes

1. In this study we generated a 6 ZF-DBD targeting an 18 bp DNA sequence (5′-AGAGATAATGGCCTAAAA-3′) in the human Gγ-globin gene promoter.
   V1: 5′-GCCCGTGTGGGTACGCTGATGTGCACGTAGGT TAGCACGTTGCGAGAACGACTTGCCACAC-3′. V2: 5′-CCTGTATGCGTCCGTTGATGTTCTGTGAGAGTGCT GTTCTGGCTGAAGGATTTACCGCAT-3′. V3: 5′-TTG TGAGTGCGCTGGTGTCTCACTAGATGTCCTGGATCAC TAAAACTCTTCCCACATTCG-3′. V4: 5′-GCCCGTGTGG GTACGCTGATGGACAGTTAGGTTGCCAGTAGTCGAGA ACGACTTGCCACAC-3′. V5: 5′-CCTGTATGCGTCCGTT GATGTCGTACTAGATTACCTGATGTGCTGAAGGATTTA CCGCAT-3′. V6: 5′-CGGTGTGAGTGCGCTGGTGTGC ACGTAGATGTGCTAGCTGACTAAAACTCTTCCCACATT CG-3′.
2. In this study we used pT7-FLAG-2 vector, which we modified by introducing additional restriction enzyme sites including *AgeI*. We used *AgeI* to express additional TG amino acids for the extended linker. Any commercially available bacterial expression vector with the desired restriction sites can be used. We use *E. coli* Stbl2 competent cells for generating ZF-DBD coding sequences.
3. This two-step sequential cloning (ZF1–3 and then ZF4–6) strategy for generating ZF-DBD coding sequences has proven to be more efficient compared to the three way ligations with vector, ZF1–3, and ZF4–6.
4. Here, we generated a 6 ZF-DBD containing a 6x His-tag at the N-terminus (Fig. 1).
5. Lowering the temperature increases solubility of the recombinant protein.
6. Alternatively, a French Press can be used to break bacterial cells.
7. We use reusable glass columns (Bio-rad) and Hislink resins (Promega). Any commercially available prepacked nickel column can be used. However, the composition of buffers may vary for different columns.
8. Any desalting method can be used for desalting. However, for downstream applications, a concentrated protein fraction may be desired. Here, we used a Vivaspin 6 concentrator for both desalting and concentrating ZF-DBDs. Typically diluting the samples ten times with lysis buffer reduces the imidazole concentration to 100 mM (which is sufficiently low for most downstream applications). ZF-DBD protein concentrations of 5–10 μM are sufficient for most applications.

9. We used band intensities to measure ZF-DBD protein concentration. Any method for measuring protein concentration can be used. However, measuring the intensity of the specific band corresponding to the ZF-DBD eliminates contributions from low amount of proteins copurifying with the ZF-DBD.
10. According to the manufacturer (Eurofins), Cy5-labeled oligonucleotides are more stable in slightly acidic buffer.
11. Increased levels of the unlabeled reverse complement DNA in the annealing reaction eliminates signals from single-stranded Cy5-labeled oligonucleotides. In our experience, additional amount of unlabeled reverse complement DNA did not affect the binding affinity of the ZF-DBD.
12. Electrophoresis at 4 °C results in sharper bands. Lower voltage may stop mobility of DNA or protein/DNA complexes during the electrophoresis. If this is the case, the voltage should be increased. On the other hand, if the voltage is too high the elevated temperature may dissociate protein/DNA complexes. Electrophoresis should be terminated when the orange dye reaches approximately 1 cm from the bottom of the gel.
13. Any imaging system that detects Cy5 signals can be used.
14. The concentration of protein necessary to observe a mobility shift will vary depending on the affinity of a particular ZF-DBD. Therefore, a range of concentrations of the ZF-DBDs should be employed that covers the entire range from no binding to complete shift of the DNA complex.
15. We used K562 cells as a representative of mammalian cells. ZF nucleases have been shown to be directly delivered to different cell types including primary cells [9].
16. In this experiment, we used serum-free medium, following the original direct delivery protocol [15]. However, according to our experience, both serum-free and serum-containing media allow efficient delivery of ZF-DBD proteins.
17. ZF-DBDs at a concentration of 100 nM were efficiently delivered to mammalian cells. The concentration may vary depending on the cell type used and the backbone of the particular ZF-DBD. The ZF-DBD used in these experiments contained the Sp1C backbone [10].
18. Heparin wash removes surface-bound proteins and reduces potential complications associated with the downstream analysis [15].
19. The NE-PER kit (Pierce) was used and yielded the desired separation of nuclear and cytoplasmic proteins. Any protocol for nuclear and cytoplasmic protein compartmentalization can be used to achieve similar results.

20. We used antibodies specific for nuclear CTCF and cytoplasmic GAPDH proteins as determinants of successful compartmentalization. Zinc finger antisera (a kind gift from Dr. Carlos Barbas, Scripps Research Institute, La Jolla, CA) used in this study recognizes the synthetic backbone (Sp1C) of the ZF-DBDs. ZF-DBDs can be cloned and expressed with desired fusion tags for detection with commercially available antibodies.

Acknowledgment

This work was supported by grants from the National Institutes of Health to J.B. (NIH, R01 DK083389, and R01 DK 052356). We are very grateful to Blanca Ostmark and Nancy Nabilsi for their help.

References

1. Klug A (2010) The discovery of zinc fingers and their applications in gene regulation and genome manipulation. Annu Rev Biochem 79:213–231. doi:10.1146/annurev-biochem-010909-095056
2. Hossain MA, Barrow JJ, Shen Y, Haq MI, Bungert J (2015) Artificial zinc finger DNA binding domains: versatile tools for genome engineering and modulation of gene expression. J Cell Biochem 116(11):2435–2444. doi:10.1002/jcb.25226
3. Gersbach CA, Gaj T, Barbas CF (2014) Synthetic zinc finger proteins: the advent of targeted gene regulation and genome modification technologies. Acc Chem Res 47 (8):2309–2318. doi:10.1021/ar500039w
4. Beerli RR, Segal DJ, Dreier B, Barbas CF (1998) Toward controlling gene expression at will: specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. Proc Natl Acad Sci U S A 95 (25):14628–14633
5. Barrow JJ, Masannat J, Bungert J (2012) Neutralizing the function of a β-globin-associated cis-regulatory DNA element using an artificial zinc finger DNA-binding domain. Proc Natl Acad Sci U S A 109(44):17948–17953. doi:10.1073/pnas.1207677109
6. Barrow JJ, Li Y, Hossain M, Huang S, Bungert J (2014) Dissecting the function of the adult β-globin downstream promoter region using an artificial zinc finger DNA-binding domain. Nucleic Acids Res 42(7):4363–4374. doi:10.1093/nar/gku107
7. Stees JR, Hossain MA, Sunose T, Kudo Y, Pardo CE, Nabilsi NH, Darst RP, Poudyal R, Igarashi K, Huang S, Kladde MP, Bungert J (2015) High fractional occupancy of a tandem maf recognition element and its role in long-range β-globin gene regulation. Mol Cell Biol 36 (2):238–250. doi:10.1128/MCB.00723-15
8. Gaj T, Liu J, Anderson KE, Sirk SJ, Barbas CF (2014) Protein delivery using Cys2-His2 zinc-finger domains. ACS Chem Biol 9 (8):1662–1667. doi:10.1021/cb500282g
9. Liu J, Gaj T, Yang Y, Wang N, Shui S, Kim S, Kanchiswamy CN, Kim JS, Barbas CF (2015) Efficient delivery of nuclease proteins for genome editing in human stem cells and primary cells. Nat Protoc 10(11):1842–1859. doi:10.1038/nprot.2015.117
10. Mandell JG, Barbas CF (2006) Zinc Finger Tools: custom DNA-binding domains for transcription factors and nucleases. Nucleic Acids Res 34(Web Server issue):W516–W523. doi:10.1093/nar/gkl209
11. Wright DA, Thibodeau-Beganny S, Sander JD, Winfrey RJ, Hirsh AS, Eichtinger M, Fu F, Porteus MH, Dobbs D, Voytas DF, Joung JK (2006) Standardized reagents and protocols for engineering zinc finger nucleases by modular assembly. Nat Protoc 1(3):1637–1652. doi:10.1038/nprot.2006.259
12. Maeder ML, Thibodeau-Beganny S, Sander JD, Voytas DF, Joung JK (2009) Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. Nat Protoc 4(10):1471–1501. doi:10.1038/nprot.2009.98

13. Thibodeau-Beganny S, Maeder ML, Joung JK (2010) Engineering single Cys2His2 zinc finger domains using a bacterial cell-based two-hybrid selection system. Methods Mol Biol 649:31–50. doi:10.1007/978-1-60761-753-2_2
14. Cathomen T, Segal DJ, Brondani V, Müller-Lerch F (2008) Generation and functional analysis of zinc finger nucleases. Methods Mol Biol 434:277–290. doi:10.1007/978-1-60327-248-3_17
15. Gaj T, Guo J, Kato Y, Sirk SJ, Barbas CF (2012) Targeted gene knockout by direct delivery of zinc-finger nuclease proteins. Nat Methods 9 (8):805–807. doi:10.1038/nmeth.2030

Chapter 28

Production, Purification, and Titration of First-Generation Adenovirus Vectors

Ramona F. Kratzer and Florian Kreppel

Abstract

Vectors based on human adenovirus are highly efficient tools for transient genetic modifications of cells or tissues in vitro and in vivo. They can be utilized for gene addition strategies, knockdown strategies and as transfer vectors for designer nucleases and CRISPR/Cas. They are characterized by high genomic stability and can be produced to high titers. This chapter describes the method how to produce, purify and titrate adenovirus vectors based on human adenovirus type 5.

Key words Adenovirus vector, First-generation vector, Transient genetic modification

1 Introduction

Adenovirus is a non-enveloped double-stranded DNA virus. Up to date more than 57 human types have been described. The DNA genome is 36–40 kB in size and the cell entry processes in vitro have been characterized to a great extent. In general, human adenoviruses exhibit only mild pathogenicity in immunocompetent individuals. However, current standard to work with adenovirus and its derived vectors requires biosafety level 2 laboratories.

Adenovirus can be converted into a replication-deficient gene transfer vector with ease [1]. Commercially available plasmid/bacmid systems contain the adenovirus genomes harboring deletions of the early gene regions *E1* and optionally *E3*. Expression of the gene products encoded by the early gene region *E1* is mandatory for the replication of the viral DNA and generation of progeny particles. Consistently, genomes with a deleted *E1* gene region cannot replicate. In place of the *E1* gene region transgene expression cassettes with heterologous promoters can be cloned. The transgene capacity of *E1*-deleted vectors is 4.5 kB, in case of *E1/E3*-deleted vectors 8 kB.

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_28,

To produce such replication-deficient vectors, producer cells that transcomplement the *E1* gene products are used. The most frequently employed cell line for this purpose is the HEK293 cell line [2]. It contains the left end of the adenovirus genome and constitutively expresses the *E1* gene region. Upon transfection of *E1*-deleted linear adenovirus vector genomes into HEK293 cells, the presence of *E1* gene products in the cells permits replication of the vector genome and the formation of vector particles. While the process of vector particle generation after transfection is not highly efficient, the few generated vector particles can be used to reinfect fresh HEK293 cells, which then produce substantial amounts of vector. This process can be repeated with increasing cell numbers 2–5 times and will yield vector preparations with $1–2 \times 10^{12}$ vector particles. The fact that adenovirus vectors can—once they have been generated as infectious vector particles—be amplified to high titers distinguishes Ad from most other viral gene transfer vector systems and is a significant advantage when continuous vector supply is required.

Vectors based on adenovirus type 5 transduce cells mainly via two receptors. The first receptor is the Coxsackie and adenovirus receptor (CAR) [3]. The vectors bind to CAR with their fiber capsid protein. Upon binding a structural change in the capsid occurs that enables binding to the secondary receptor: $\alpha_v\beta_{3/5}$ integrins [4]. Binding to the latter mediates uptake of the vector particles via receptor-mediated endocytosis. Subsequently, the particles evade from the early endosome and are transported to the nucleus. Here, the vector genome is translocated into the nucleus by the use of cellular mechanisms [5]. The overall process is fast (20–40 min) and highly efficient. Since the receptors CAR and integrins are expressed on many cell types, a wide variety of cells of different origin can be transduced.

Overall *E1/E3*-deleted adenovirus vectors are characterized by the following features:

- A large transgene capacity of 8 kB.
- High genomic stability.
- Ability to infect a wide variety of quiescent and proliferating cells.
- Lack of integration into the host cell genome.
- Production to high titers.

Here, we describe the generation and characterization of adenovirus vector particles based on Ad5. This is the best characterized and most versatile vector system available up to date. Cloning of transgene expression cassettes into the Ad plasmid/bacmid systems is not subject of this chapter since it can be achieved by conventional standard techniques. Instead, we focus on the processes to generate infectious vectors from the plasmid material.

2 Materials

Prepare all buffers and solutions in ultrapure and optionally autoclaved, double-distilled water (ddH_2O) and use analytical grade chemicals. Sterilize buffers by filtration (0.45 μm mesh pore size) or autoclaving (20 min at 121 °C), as indicated. Store buffers at room temperature if not indicated otherwise.

2.1 Adenovirus Genome Transfection

1. TE buffer: 10 mM Tris, 1 mM EDTA, pH 8.5. Autoclave.
2. Buffer-saturated phenol.
3. Chloroform/isoamyl alcohol (24:1): 24 parts of chloroform, one part of isoamyl alcohol.
4. Sodium acetate: 3 M sodium acetate, pH 5.2. Autoclave.
5. Ethanol absolute.
6. 70% Ethanol: Ethanol absolute, diluted to 70%.
7. Dulbecco's phosphate buffered saline (DPBS): commercial cell culture grade buffer.
8. Polyethylenimine (PEI) solution: linear 22 kDa PEI, 7.5 mM, pH 7.0. Sterile filtrate. Store at 4 °C.
9. NaCl solution: 150 mM NaCl. Autoclave.
10. Ad producer cells: *E1*-complementing cell line for production of first generation Δ*E1* Ad vectors, e.g., HEK293 [2] (ATCC, CRL-1573).

2.2 Adenovirus Harvest, Rescue, and Amplification

1. Ad buffer: 50 mM HEPES, 150 mM NaCl, pH 7.6. Sterile filtrate. Prepare fresh, protect from light and store at 4 °C.
2. Culture plates, 6 cm and 15 cm diameter.
3. Cell scraper.
4. Conical 50 ml centrifugation tubes, e.g., BD Falcon.
5. Liquid nitrogen.
6. Water bath at 37 °C.

2.3 Adenovirus Purification

1. Liquid nitrogen.
2. Water bath at 37 °C.
3. Ad buffer: 50 mM HEPES, 150 mM NaCl, pH 7.6. Sterile filtrate. Prepare fresh, protect from light, and store at 4 °C.
4. CsCl step gradient buffer: $\rho_{CsCl} = 1.27$ g/cm^3 in Ad buffer, pH 7.6: weigh 18.47 g CsCl and fill up to 50 ml with Ad buffer. Sterile filtrate. Prepare fresh, check density by weighing 1 ml, protect from light, and store at 4 °C.
5. CsCl step gradient buffer: $\rho_{CsCl} = 1.41$ g/cm^3 in Ad buffer, pH 7.6: weigh 27.42 g CsCl and fill up to 50 ml with Ad buffer.

Sterile filtrate. Prepare fresh, check density by weighing 1 ml, protect from light, and store at 4 °C.

6. CsCl continuous gradient buffer: $\rho_{CsCl} = 1.34$ g/cm^3 in Ad buffer, pH 7.6: weigh 22.71 g CsCl and fill up to 50 ml with Ad buffer. Sterile filtrate. Prepare fresh, check density by weighing 1 ml, protect from light, and store at 4 °C.
7. Glycerol. Autoclave.
8. 200 ml centrifugation tubes, e.g., NUNC.
9. 13.2 ml ultracentrifugation tubes, e.g., 13.2 ml UltraClear, Beckman Coulter.
10. Ultracentrifuge with suitable rotor, e.g., Sorvall Discovery 90SE with TH-641 rotor or Beckman SW41.
11. Syringes and needles.
12. PD-10 size exclusion chromatography column, GE Healthcare.
13. Tripod with clamps for ultracentrifugation tubes.
14. Optional: Gooseneck lamp.

2.4 Adenovirus Titration by Measurement of OD_{260}

1. Sodium dodecyl sulfate (SDS) solution: 10% SDS in autoclaved ddH_2O.
2. Dulbecco's phosphate buffered saline (DPBS): commercial cell culture grade buffer.
3. Blank sample: Ad buffer, 10% glycerol.

2.5 Adenovirus Titration by Radioactive Probing of DNA (Slot Blot)

1. ddH_2O. Autoclave.
2. NaOH solution: 0.8 M NaOH. Autoclave.
3. EDTA solution: 50 mM EDTA in DPBS. Autoclave.
4. SSC buffer (20×): 3 M NaCl, 0.3 M tri-sodium citrate, pH 7.0. Autoclave. Use SSC buffer (20×) to prepare SSC working solution (2×).
5. Dextran sulfate solution: 50% dextran sulfate. Dissolve under heating and constant stirring.
6. TE buffer: 10 mM Tris, 1 mM EDTA, pH 7.5. Autoclave.
7. Salm sperm DNA: 10 mg/ml in TE buffer, pH 8.5. Dissolve by vortexing and heating to 37 °C. Store at 4 °C.
8. Sodium dodecyl sulfate (SDS) solution: 10% SDS in autoclaved ddH_2O.
9. Milk solution: 5% milk powder in SDS solution.
10. Hybridization buffer: 2 ml SSC buffer, 2 ml Milk solution, 1 ml Salmon sperm DNA, 4 ml Dextran sulfate solution, 11 ml autoclaved ddH_2O. Vortex, boil for 10 min, and cool down on ice.

11. Wash buffer I: 100 ml SSC buffer, 10 ml SDS solution, 890 ml autoclaved ddH_2O.
12. Wash buffer II: 5 ml SSC buffer, 10 ml SDS solution, 985 ml autoclaved ddH_2O.
13. Positively charged 0.45 μm Nylon membrane, e.g., Biodyne B, Pall Corporation.
14. Slot blot manifold, e.g., PR-648, Hoefer.
15. Vacuum pump.
16. Rediprime II DNA Labeling System master mix, Amersham.
17. ^{32}P-labelled α-dCTP, 50 μCi.
18. Phosphoscreen cassette, e.g., Storage Phosphor Screen, GE.
19. Phosphoscreen reader, e.g., Phosphoimager BAS 1000, Fuji with Aida software.
20. Oven at 120 °C.
21. Oven at 68 °C with unit for rotation of ≈5 cm diameter tubes.
22. Two heat blocks for 1.5 ml tubes at 37 and 98 °C.
23. Two big hubs to soak the Slot blot manifold.
24. Cylindrical glass tube, ≈5 cm diameter.
25. Standard DNA: Prepare plasmid DNA of a Δ*E1* Ad vector construct, and bring it to a concentration of 2×10^6 copies/μl.
26. DNA probe for radioactive labelling: Prior to titration, you have to design a DNA probe which is suitable for your construct. To titrate Δ*E1* Ad5 titers, the probe can be directed to capsomer sequences, e.g., fiber or hexon. An Ad5 fiber probe can be obtained by PCR (5′–3′primer sequences ATGAAGCGCGCAAGACCGTCTG and CCAGATATTG GAGCCAAACTGCC).

3 Methods

3.1 Adenovirus Genome Transfection

1. The day before PEI transfection, seed 10^6 HEK293 cells per 6 cm cell culture plate.
2. Prior transfection, cloned Δ*E1* Ad5 genomes have to be released from the bacterial plasmid/bacmid backbone by restriction digest of circular plasmids/bacmids with a suitable restriction enzyme to enable replication of the Ad genome in eukaryotic producer cells. This restriction digest is performed by digesting 10–25 μg DNA in a total volume of 100 μl as a 10- to 50-fold overdigestion (*see* **Note 1**).

3. Purify linearized DNA by phenol/chloroform extraction. Bring digestion reaction mixture to 200 µl with TE buffer, add 200 µl of phenol. Mix vigorously and centrifuge for 10 min, 20,000 × *g*. Collect the aqueous (typically: upper) phase and add 200 µl of chloroform/isoamyl alcohol. Mix vigorously and centrifuge for 10 min, 20,000 × *g*. Collect the upper phase, add 500 µl of Ethanol absolute and 20 µl of 3 M sodium acetate. Precipitate for 15 min, 4 °C, 20,000 × *g*, wash with 70% EtOH, and resuspend in 30 µl of TE buffer. Determine DNA concentration by OD_{260}.
4. Bring 5 µg of linearized Δ*E1* Ad5 DNA to a volume of 250 µl with NaCl solution. Bring 60 µl of PEI solution to a volume of 250 µl with NaCl solution. Prepare PEI/DNA complexes by rapid addition of PEI into DNA. Incubate mixtures for 10 min, RT.
5. Wash Ad producer cells with DPBS, replace medium, and transfect cells by dropwise addition of PEI/DNA complexes. Exchange medium after 3 h to O/N incubation and monitor Ad producer cells daily for signs of cytopathic effect (CPE) (*see* **Note 2**).

3.2 Adenovirus Harvest, Rescue, and Amplification

1. At 7–12 days after PEI transfection, Ad producer cells will show full CPE (*see* Fig. 1). Harvest cells by scraping plates and transferring culture supernatants to a 50 ml conical centrifugation tube (*see* **Note 3**).
2. After centrifugation for 10 min, 4 °C, 400 × *g*, resuspend the pellet in 2 ml of Ad buffer, and rescue the vector through three cycles of freeze/thaw in liquid nitrogen and 37 °C water bath.

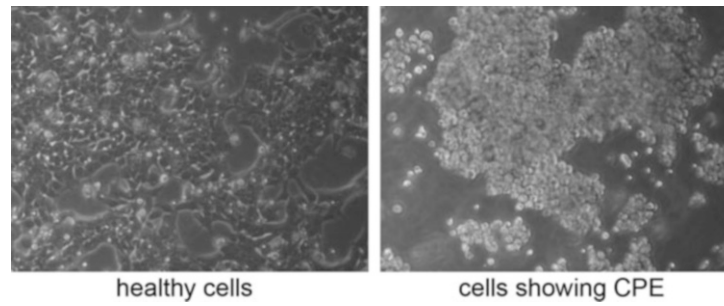


Fig. 1 Adenovirus producer cells showing CPE. The cytopathic effect is an indicator for adenovirus virus vector production in E1-transcomplementing producer cells. Non-infected cells (*left panel*) are attached to the dish surface. Infected cells (*right panel*) round up and start to detach from the dishes in berry-like structures. This is the optimal timepoint to harvest the cells since all virions will still be contained in the cells and, thus, medium supernatant can be discarded

3. The lysate containing rescued vector particles is used to reinfect about 1–2 × 10^7 Ad producer cells seeded the day before on one to two 15 cm culture plates (*see* **Note 4**).
4. At 48 h after the first reinfection, Ad producer cells should show full CPE. Harvest cells for the final amplification which is done by reinfection of about 1–2 × 10^8 cells (ten to fifteen 15 cm culture plates). For harvest of cells and rescue of vector particles, proceed as described in **step 1** and **step 2** (*see* **Note 5**).
5. At 48 h after the final reinfection, harvest cells for purification of Ad vector particles. Continue with the protocol in Subheading 3.3.

3.3 Adenovirus Purification

1. To harvest the cells of the final amplification, transfer cells and supernatants to 200 ml centrifugation tubes as described before.
2. After centrifugation for 10 min, 4 °C, 400 × *g*, resuspend the pellet in 3 ml of Ad buffer, and rescue the vector through three cycles of freeze/thaw in liquid nitrogen and 37 °C water bath.
3. After centrifugation for 10 min, 4 °C, 5000 × *g*, load the supernatant (i.e., cell lysate containing rescued vector particles) on top of a CsCl step gradient (upper phase: 5 ml of ρ_{CsCl} = 1.27 g/cm^3, lower phase: 3 ml of ρ_{CsCl} = 1.41 g/cm^3) and fill the tube to the top. Adjust weight of opposite tubes with Ad buffer (to 0.0 g difference), and ultracentrifuge for 2 h, 4 °C, 176,000 × *g* (*see* **Note 6**).
4. Fix the ultracentrifugation tube (with tripod and clamp), place gooseneck lamp above the tube, clean the tube's surface with Ethanol, and collect the vector band by puncturing the tube with a syringe, and aspiring the band (*see* Fig. 2 and **Note 7**).
5. Transfer the collected band to a fresh ultracentrifugation tube, mix with ρ_{CsCl} = 1.34 g/cm^3, and fill the tube to the top. Adjust weight of opposite tubes with ρ_{CsCl} = 1.34 g/cm^3, and ultracentrifuge for 20–24 h, 4 °C, 176,000 × *g*.
6. Equilibrate a PD-10 column with five times 5 ml of Ad buffer.
7. Collect vector band as described in **step 4**. Bring to a final volume of 2.5 ml with Ad buffer, and load onto the prepared PD-10 column. Discard flowthrough (*see* **Note 8**).
8. Add 5 ml of Ad buffer and collect the eluate in fractions of 1 ml each (*see* **Note 9**).
9. Combine the second and third fraction, add glycerol to a final concentration of 10%, aliquot in suitable volumes, and store at −80 °C (*see* **Note 10**).

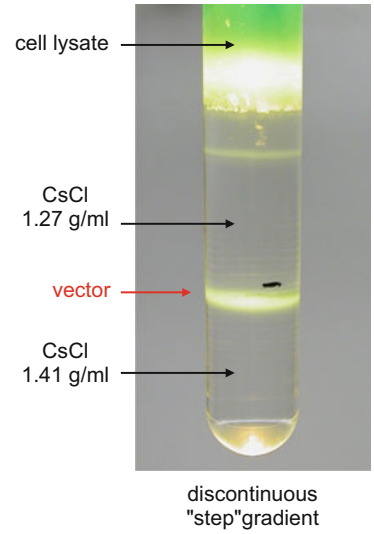


Fig. 2 Example of a discontinuous CsCl gradient for purification of Ad vectors. The photograph shows the result of the first discontinuous gradient centrifugation. The *thick band* in the lower part of the tube contains the vector. The *thin band* above consists of incomplete particles and can be discarded. The *upper part* contains the cell debris. Note that the *green color* is EGFP. The vector produced shown here contained an EGFP expression cassette and substantial amounts of EGFP are synthesized during vector production

3.4 Adenovirus Titration by Measurement of OD_{260}

1. Denaturate vector capsids and release vector genomes to measure OD_{260}. Use 20 µl of purified vector (obtained in Subheading 3.3, **step 9**), add 79 µl of DPBS and 1 µl of SDS solution. For the blank measurement, use 20 µl of blank sample instead of purified vector.
2. Vortex mixtures, incubate for 10 min, 56 °C, and determine OD_{260}.
3. Physical vector titer as vector particles per microliter [vp/µl] can be calculated by multiplying the measured OD_{260} value, the dilution factor, and the empirically determined extinction coefficient for Ad (1.1×10^9 vp equal one OD_{260} unit per µl [6, 7]) (*see* **Note 11**).

3.5 Adenovirus Titration by Radioactive Probing of DNA (Slot Blot)

1. To determine the physical vector titers by radioactive probing (Slot Blot [8]), soak Slot blot manifold O/N in ddH_2O. Prior to use, soak the manifold for 20–30 min in 1:2 diluted NaOH solution, and thoroughly rinse all slots with desalted water (*see* **Note 12**).
2. To prepare the standard samples, add 1, 5, 10, 20, 50 µl of standard DNA (in duplicates) to 200 µl of EDTA solution (*see* **Note 13**).

3. Prepare a 1:100–1:200 dilution of purified vector in DPBS. To prepare the vector samples, add 2, 10, 20 μl of vector dilution (in duplicates) to 200 μl of EDTA solution.
4. Add 200 μl of NaOH solution to each tube, vortex, and incubate for 20–60 min, RT.
5. Equilibrate membrane (11.2 cm × 8.1 cm) in 1:2 diluted NaOH solution, mount into Slot blot manifold, and connect to vacuum pump. Apply vacuum for 2–5 min.
6. Thoroughly vortex all standard and vector samples prior to loading of 300 μl into the slots.
7. Apply vacuum until all slots are emptied. Remove the membrane, rinse it in SSC working solution, and bake for 20 min, 120 °C to crosslink DNA (*see* **Note 14**).
8. Transfer the membranes to cylindrical glass tubes, and block with hybridization buffer for 1 h, 68 °C under constant rotation.
9. To prepare the radioactively labelled DNA probe for hybridization, denaturate 100 ng of DNA probe (e.g., Ad5 fiber probe) in 45 μl TE, pH 7.5 by boiling for 5 min, and cool down on ice (*see* **Note 15**).
10. Add denatured DNA probe and 50 μCi of ^{32}P-labelled α-dCTP to the labelling master mix, and incubate in a heat block for 30 min, 37 °C. Boil for 2 min, 98 °C, and add labelling reaction mixture into the glass tube with the blocked membrane and hybridize O/N at 68 °C under constant rotation.
11. Discard radioactive hybridization buffer (*see* **Note 16**).
12. Rinse once with wash buffer I, and wash twice for 10 min, 68 °C under constant rotation. Rinse once with wash buffer II, wash once for 10 min, 68 °C and once for 5 min, 68 °C under constant rotation.
13. Remove the membranes from the glass tube, rinse in wash buffer II, air-dry, and place the membranes in a Phosphoscreen cassette.
14. Read out the radiation signal with a Phosphoscreen reader after 2 h at earliest, and evaluate radioactive signal intensity with appropriate software. Calculate titers of vectors samples according to the standard DNA (*see* Fig. 3).

4 Notes

1. Select a suitable restriction endonuclease according to sequence analysis of your circular construct. You can select any enzyme that shows at least one restriction site outside the Δ*E1* Ad5 genome sequence and will result in linearization of

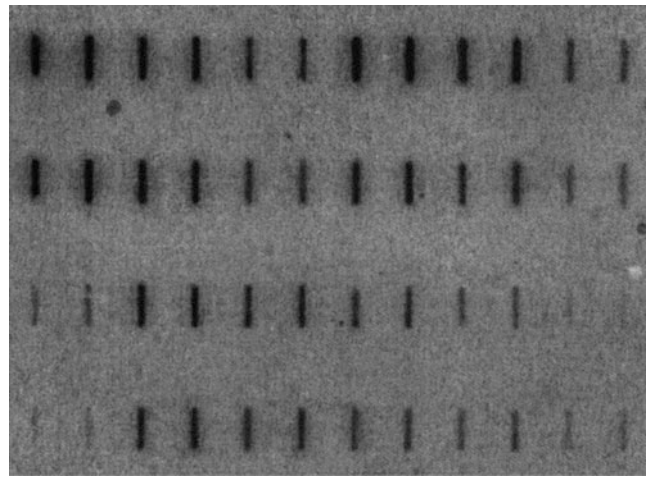


Fig. 3 Example of a vector titration by the Slot-blot procedure. Signal intensities can be quantified using a phosphoimager and appropriate software. The *upper two lanes* contain standard DNA and the *lower two lanes* the vector preparations

your circular construct. Typically, commercial plasmid/bacmid systems are equipped with suitable restriction sites.

2. Cytopathic effect (CPE) of Ad producer cells is caused by Ad proteins that are expressed during the infectious cycle. The cytopathic effect is characterized by rounded cells which detach in berry-like clusters (*see* Fig. 1).
3. The time point of full CPE and harvest will depend on your transfection efficiency, and the replication efficiency of your vector construct. For freeze/thaw, take care to use tubes that do not burst. Note that the appearance of the CPE may differ from vector to vector since transgenes products can contribute to toxicity.
4. After freeze/thaw, the lysate does not need to be cleared from cell debris before reinfection of Ad producer cells.
5. Prior to the final amplification, you should titrate the vector-containing lysate to reach optimum CPE at 48 h after reinfection. For that, seed five 6 cm plates of Ad producer cells, and reinfect the cells with varying amounts of lysate (1, 2, 10, 20, 50 μl). According to the observed CPEs at 48 h, decide on the ideal amount of lysate and calculate the required volume to be used per 15 cm plate.
6. Be gentle when preparing the step gradients. You will be able to see the phase border between the CsCl solutions of different densities. The vector band after the first step gradient ultracentrifugation should run at the phase border. Make sure you only load the supernatant onto the gradient. If you transfer parts of

the cell debris, this will complicate aspiration of the vector band after ultracentrifugation.

7. After the first step gradient, you should only observe one clear band at the phase border. You might also observe blurry structures if you unintenionally transferred cell debris. The discontinuous gradient removes up to 90% of contaminating cellular debris and concentrates the vector particles.
8. After the second continuous gradient, only one band should occur. If multiple bands occur vector particles with different genome sizes may have been produced. In that case try to separately collect the different bands, prepare DNA, and analyze the genomes by restriction analysis.
9. Use 1.5 ml tubes and mark the 1.0 ml filling level. This will help to easily collect 1 ml fractions.
10. The second and third fraction contain the highest concentration and highest purity and concentration of vector particles.
11. The formula to calculate the physical vector titer, i.e., vector particle concentration per microliter is OD_{260} value $\times$ 5 $\times$ 1.1 $\times$ 10^9 [vp/µl].
12. 1:2 diluted NaOH solution (0.4 M NaOH) is prepared from autoclaved 0.8 M NaOH and ddH_2O. Make sure you keep some 0.8 M NaOH for **step 4** of Subheading 3.5.
13. This equals to a standard ranging from 2 $\times$ 10^6 to 1 $\times$ 10^8 copies.
14. If you wish, you can store the baked membrane at RT, and continue with **step 8** of Subheading 3.4 on another day. If you wish to titrate infectious vector particles, a reference cell line (such as A549 cells, ATCC, CCL-185) is transduced, and cells are harvested with EDTA solution. Lysis, denaturation, and slot blotting are done analogously to titration of physical vector titers.
15. Make sure that all steps in Subheading 3.5 following **step 9** are carried out in appropriate laboratories ensuring work safety.
16. Consider appropriate disposal of radioactive waste.

References

1. Volpers C, Kochanek S (2004) Adenoviral vectors for gene transfer and therapy. J Gene Med 6 (Suppl 1):S164–S171. doi:10.1002/jgm.496
2. Graham FL, Smiley J, Russell WC, Nairn R (1977) Characteristics of a human cell line transformed by DNA from human adenovirus type 5. J Gen Virol 36:59–74. doi:10.1099/0022-1317-36-1-59
3. Bergelson JM, Cunningham JA, Droguett G et al (1997) Isolation of a common receptor for Coxsackie B viruses and adenoviruses 2 and 5. Science 275:1320–1323
4. Wickham TJ, Mathias P, Cheresh DA, Nemerow GR (1993) Integrins alpha v beta 3 and alpha v beta 5 promote adenovirus internalization but not virus attachment. Cell 73:309–319

5. Trotman LC, Mosberger N, Fornerod M et al (2001) Import of adenovirus DNA involves the nuclear pore complex receptor CAN/Nup214 and histone H1. Nat Cell Biol 3:1092–1100. doi:10.1038/ncb1201-1092
6. Maizel JV, White DO, Scharff MD (1968) The polypeptides of adenovirus: I. Evidence for multiple protein components in the virion and a comparison of types 2, 7A, and 12. Virology 36:115–125
7. Mittereder N, March KL, Trapnell BC (1996) Evaluation of the concentration and bioactivity of adenovirus vectors for gene therapy. J Virol 70:7498–7509
8. Kreppel F, Biermann V, Kochanek S, Schiedner G (2002) A DNA-based method to assay total and infectious particle contents and helper virus contamination in high-capacity adenoviral vector preparations. Hum Gene Ther 13:1151–1156. doi:10.1089/104303402320138934

INDEX

A

Adenovirus 377–384
Adenovirus vector 377–387
Agrobacterium tumefaciens 341
Algorithm 3–9, 12, 23–26, 29, 34, 42, 43, 48, 50, 66, 68, 159, 182, 200, 206, 357
Annealing 41, 128, 147, 154–156, 169, 192, 229, 233, 326, 332, 373
Annotation 7, 22–24, 55–58, 60–63, 66–68, 71, 200, 201, 206–207
Antibody array 261–269
Antigen array 271–276
Argonaute 20, 180
Autoantibody 271
Axon 231–241

B

Backbone 5, 6, 47, 50, 169, 344, 364–366, 373, 374, 381
Barcoded particle 210
Bean pod mottle virus (BPMV) 311–318
Binding affinity 153, 161, 362, 373
Biomarker 179, 209, 210, 292
Biotinylated 211, 261, 263, 265, 267, 282, 285, 286, 332, 333

C

Canonical base pair 11, 12
Cas9 ..... 165–175, 323
Catalyzed signal amplification (CSA) 280, 282, 285
CDNA libraries 179–194, 198
Cell lysis 168, 280, 282, 353, 355–357
Cell type-specific profiling 102
Chromatin immuno precipitation (ChIP) 99
Chromatin interactions 79
Chromatin proteins 99–121
Chromatin remodeling 19
Chromatin structure 81, 99, 120
Chromosome conformation capture (3C) 80, 81
Classifier 24, 30, 34, 35
Clusters of orthologous groups (COGs) 60, 61, 68
Command line 31, 42, 48, 93, 97
Comparative modeling 39–42, 50
Compartmentalized culture 232–236
Competing endogenous RNA (ceRNA) 18, 24–27
Conformational space 4
Convergent evolution 49
Copy number variation (CNV) 135–148
Coregulation 19
CpG methylation 125, 126, 130, 133, 322
CRISPR/Cas9 165–175
Cross-linking 41, 80, 87, 99
Cycle threshold (CT) 135
Cytokine 292
Cytosines preceding a guanine (CpG) 125, 126, 130, 133, 321, 322, 329

D

Dam identification (DamID) 99–121
Data processing 118–119
Density map 46
Dicer 18, 180
Differential centrifugation 291
Direct protein delivery 361–374
Dissociation constant (K_d) 156, 159, 162, 370
DNA
- demethylation 321–333
- isolation 42, 110, 171, 175
- methylation 100, 101, 103, 106, 111, 121, 125–133, 322, 324–326, 328–330, 332
- replication 99, 100

DNA methyltransferase protein (DNMTs) 321, 322
DNA–protein interaction 151, 152
Docking 46
Double strand breaks (DSB) 165, 166
Droplet digital PCR (ddPCR) 135–148
Drosophila melanogaster 100
Dynamic programming 42, 43

E

Electrophoresis 83, 91, 132, 168, 172, 181, 182, 184, 186, 188, 190–194, 198, 207, 233, 238, 246, 303, 341, 353, 363, 367, 369, 373
Electrophoretic mobility shift 365, 368–370
Encoded particle 222, 227
Encoded primer-immobilized networks (PIN) 221, 222
Energy profile 44–46

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9, © Springer Science+Business Media LLC 2017

Engineered zinc finger DNA-binding domain 361–374
Epigenetic editing 322–324
Epigenetic modification 23–24, 166
Epigenomics 180
Escherichia coli 59, 317, 341, 343
Exosome 217, 218, 291–306
Expression profile 21–24, 179–194

F

Fasta file 30–33, 97, 130
First-generation vector 377–387
FLP-inducible DamID 99–121
"Flp-out" approach 102
Fluorescence spectroscopy 41
Fluorescent proteins 166
Fold assignment 40–43, 50
Force field 4, 40, 41, 51
Free modeling 39

G

Garden pea 312
Gene
  expression 18, 19, 22, 29, 56, 120, 135, 179–181, 197, 198, 311–318, 323, 324, 346, 352, 356, 362, 378
  silencing 316, 351, 352, 357–359
  tagging 165–175
Genome
  organization 79–97
  sequencing 39
Genome-wide binding patterns 99
Genomic(s) 17, 19, 22–23, 30, 42, 55, 56, 61, 68–70, 80, 81, 100, 101, 103, 107–115, 119, 120, 126–128, 130–132, 166, 171, 172, 175, 180, 181, 193, 312, 321–333, 361, 378
  binding sites 100, 101
  context 61, 68–69
Genotyping 138, 139
Glycosylation 261
Gradient 41, 152, 153, 158, 162, 240, 246, 252, 253, 259, 264, 274, 283–285, 300, 301, 379, 380, 383, 384, 386, 387
gRNA 165–175

H

Hairpin 4, 6, 8, 11, 12, 18, 29, 30, 33, 180, 197
Haptoglobin 137
Hi-C 79–97
Hidden Markov model (HMM) 63, 65, 66
High throughput 17, 18, 20, 27, 55, 68, 100, 107, 117–119, 130, 136, 182, 198, 210, 255, 261, 279, 293, 312
Homology 40, 41, 56–57, 59–63,65, 197, 322
Hybridization 100, 107, 210, 211, 213, 215, 218, 380, 385
Hydrogel particles 210

I

ICAM-1 promoter 324
Illumina sequencing 118, 181
In-gel digestion 255, 256
In-nucleus ligation 81
In-solution ligation 81
Isoelectrofocusing 248–249, 251, 252

K

Knockdown 18, 22, 102, 337–348, 351–359

L

Legume 312, 313
Ligation 80–82, 86, 89, 93, 105, 114, 115, 118, 167, 169, 170, 182, 183, 187, 191, 194, 198, 202, 203, 344, 348, 372
Lipoplexes 352, 358
Liquid biopsy 228
Long noncoding RNA (lncRNA) 17–19, 21–27, 33, 179, 232
Long-read sequencing 125–133
Loop entropy 5, 6, 9, 12

M

Machine learning 30, 61
Mammalian cells 351, 361–374
Mapping 96, 97, 99–121, 200
Mass spectrometry 70, 250, 291–306
Methylase 166
Microarray 23, 61, 69, 70, 100, 107, 181, 197, 221, 231, 263, 264, 267, 268, 271, 273, 274, 276, 282, 284, 287, 331
MicroRNA (miRNA) 17–26, 29–36, 179–181, 184, 185, 197–218, 221–229, 337–348
Minimum Euclidean distance 34
Minimum free energy (MFI) 4, 6, 13, 31–33
miRNA. *See* MicroRNA (miRNA)
miRNA annotation 201
miRNA expression analysis 221
miRNA precursor 29, 30, 35
miRNA prediction 33, 35
miRNA profiling 206–210, 221

miRNome 180, 181
Mitochondrial DNA 32, 117
Molecular dynamics simulation 4
Moonlighting 58, 59
Motoneurons 232, 234
Multiple sequence alignments (MSA) 65
Multiplex circulating miRNA assay 209
Multiplex DNA methylation analysis 125
Multiplex miRNA profiling 209–218
Multiplex qPCR 221–228
Multiplex sequencing 179–194

N

Near infrared (NIR) fluorescence 280, 285
Next-generation sequencing 85
Non-coding RNAs 30, 33
Non-homologous end joining 165–166
Normalization 49, 185, 198, 200, 201, 206, 245, 269, 306
Nuclear magnetic resonance (NMR) spectroscopy 39, 41

O

Off-target effects 358
Oligonucleotides 84, 85, 94, 106, 139, 154–156, 167, 169, 173, 186, 352–354, 356, 358, 364–366, 373
Orthologs 56, 61, 63, 71, 166

P

Paralogs 56, 61, 63
Phaseolus vulgaris 312–317
Phosphorylation 85, 169
Phylogenetics 56
PIN. *See* Encoded primer-immobilized networks (PIN)
Pisum sativum 311–318
Polymerase chain reaction (PCR) 80, 100, 125, 135–148, 167, 182, 198, 210, 221–229, 232, 326, 341, 362, 381
Post-transcriptional gene silencing (PTGS) 312
Post-translational modification 255
Primer 83–85, 90, 91, 95, 100, 101, 105, 108, 114–116, 118, 126, 128, 129, 132, 139–141, 145, 146, 167–169, 174, 175, 187, 191, 192, 194, 198, 200, 202–205, 207, 208, 211, 214, 221, 222, 224–229, 233, 236, 237, 241, 326, 329, 330, 333, 340–343, 345–347, 364–365, 367
Probability density 41
Protein
  array 261–269, 271–276, 279–288
  digestion 258, 294–295, 298–300
  domain 58, 60, 61, 71, 361
  expression 56, 102, 135, 209, 267, 279, 287, 324, 351–359, 362
  families 60, 63–65
  function 55–72, 135
  labeling 265
  purification 363
  sequence 40, 41, 56, 57, 61–63, 65–68
  structure 39–52, 55, 56, 65, 66, 138
Protein–protein interaction 56, 59, 61, 69, 70, 255
Proteome 60, 62, 63, 68, 250, 292, 293, 302
Proteomics 70, 180, 245–253, 255–259, 287, 292
Protospacer adjacent motif (PAM) 165, 169, 174
Pseudoknot 6, 8
Pyrosequencing 125, 326, 329–332
Python 42, 44, 45, 48

Q

qPCR. *See* Quantitative PCR (qPCR)
QRT-PCR 324
Quantitation cycle (Cq) 135
Quantitative PCR (qPCR) 91, 107, 135, 136, 205, 221–229, 236, 238, 356

R

Recombinant protein purification 362
Replication-deficient gene transfer vector 377
Reverse-phase protein array (RPPA) 279–288
Reverse transcriptase 183, 187, 192, 202, 205, 236, 359
Reverse transcription qPCR (RT-qPCR) 181, 197–208
Riboswitch 4, 10, 12, 13
Rice 337–348
RNA
  function 17–27
  sequencing 3, 27, 197–208
  structure 3–13
RNAi. *See* RNA interference (RNAi)
RNA interference (RNAi) 179, 351–359
Root mean square deviation (RMSD) 8, 11–13, 45, 47, 51
RT-qPCR. *See* Reverse transcription qPCR (RT-qPCR)

S

Saccharomyces cerevisiae 59
Secondary antibody 272, 282, 364
Secondary structure 10, 13, 29, 33, 41, 43, 66, 67, 146, 190
Short interfering RNA (siRNAs) 337, 351–354, 356–358

Short tandem target mimic (STTM) 337–348
Single-molecule real-time bisulfite sequencing 125, 133
Single molecule real-time bisulfite sequencing (SMRT-BS) 125–133
siRNAs. *See* Short interfering RNA (siRNAs)
Site directed mutagenesis 41, 42
Small noncoding RNA..... 29, 179, 197
Stem-loop 29, 200, 227, 338, 342
Streptavidin..... 83, 87, 261, 263, 266, 282, 286, 288, 326, 330
Structure modeling 39–52
Support vector machines (SVM)..... 66, 68
Synthetic siRNA 352

T

Taq DNA polymerase 348
T7E1 assay..... 167, 172, 175
Ten-eleven translocation 2 321–333
Thermophoresis 151–163
Third generation sequencing 126
Toxicity 356–359, 386
Transcription 22, 23, 99, 103, 151, 152, 166, 312, 322, 323
  factors..... 27, 153, 160, 166, 322
  termination 151, 152
Transcription termination factor 1 ((TTF-I)..... 151, 153–160
Transcriptome profiling 231–241
Transduction 209, 234, 325–328
Transfection 170–171, 174, 327, 331, 352–354, 356–358, 378, 379, 381–382, 386
Transient genetic modifications 377
Transmission electron microscopy (TEM) 303
Tryptophan fluorescence 293, 294, 297–298, 303, 304

U

Ultracentrifugation 380, 383, 387
Ultrafiltration 292
Ultra performance liquid chromatography 255

V

Vector..... 34, 66, 106–108, 120, 167, 169–170, 173, 174, 311–318, 325, 326, 340, 342–348, 364–367, 372, 377–387
Virus-induced gene silencing (VIGS)..... 311–318

W

Western blot 268, 279, 297, 303

X

X-ray crystallography 3

Z

Zinc finger 166, 323–327, 331, 332, 361–374