


Methods in
Molecular Biology 1712

Springer Protocols



Steven R. Head
Phillip Ordoukhanian
Daniel R. Salomon *Editors*

Next Generation Sequencing

Methods and Protocols

 Humana Press

Download From: AghaLibrary.com

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB,

Next Generation Sequencing

Methods and Protocols

Editors

Steven R. Head

*Next Generation Sequencing and Microarray Core Facility,
The Scripps Research Institute,
La Jolla, CA, USA*

Phillip Ordoukhanian

*Next Generation Sequencing and Microarray Core Facility,
The Scripps Research Institute,
La Jolla, CA, USA*

Daniel R. Salomon

*Department of Molecular Medicine, The Scripps Research Institute,
Center for Organ and Cell Transplantation,
La Jolla, CA, USA*

 **Humana Press**

Editors

Steven R. Head
Next Generation Sequencing and Microarray
Core Facility
The Scripps Research Institute
La Jolla, CA, USA

Phillip Ordoukhanian
Next Generation Sequencing and Microarray
Core Facility
The Scripps Research Institute
La Jolla, CA, USA

Daniel R. Salomon
Department of Molecular Medicine
The Scripps Research Institute
Center for Organ and Cell Transplantation
La Jolla, CA, USA

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-7512-9 ISBN 978-1-4939-7514-3 (eBook)
<https://doi.org/10.1007/978-1-4939-7514-3>

Library of Congress Control Number: 2017960416

© Springer Science+Business Media, LLC 2018

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

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

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

Printed on acid-free paper

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

Dedication

This volume is dedicated to the memory of our friend, colleague, mentor, and surf buddy, Daniel R. Salomon, M.D.

Preface

The revolution in high-throughput sequencing technology—Next Generation Sequencing (NGS)—has transformed the science of genomics in the last decade. The resulting deluge of new methods, protocols, and techniques to answer fundamental questions in biology has enabled highly efficient strategies for addressing problems of DNA, RNA, and their interactions with proteins. Each chapter in this *Methods in Molecular Biology* book describes a current state of the art in NGS application and is intended as a resource for researchers with all levels of NGS experience who wish to expand their knowledge and practical skills in high-throughput sequencing. Examples of the challenges of the role of NGS in basic research protocols data analysis, as well as clinical applications are included.

This book covers a wide range of various fields of research, with the common thread being NGS-related methods and applications, as well as some analysis and interpretation of the data obtained. The first two chapters focus on the highly dynamic processes of translational and transcriptional profiling of a cell. In the first chapter, polysome and ribosome isolation by sucrose gradient is used to investigate translational activity, both approaches are described, and the information obtained from each is discussed. In the second chapter, isolation of cell nuclei and the nascent RNA transcripts therein allow a look at the nascent transcriptome of a cell. The third chapter focuses on a method to detect copy number alterations (CNAs) using whole genome amplification and low pass whole genome sequencing. Chapters 4 and 5 touch on more targeted sequencing applications. The first uses small “bait” oligonucleotides to “fish” out long pieces of DNA (~2 Kb) from the genome combined with long read sequencing using the MinION (Oxford Nanopore, Inc.) allowing for the interrogation of these unknown flanking regions not contained in the baits. Chapter 5 deals with an extremely efficient and cost-effective method called “Hi-Plex” to characterize known polymorphic loci using a highly multiplexed amplicon-based approach to look for genetic variants.

Chapter 15 focuses on DNA structural rearrangements using a method called, “Hi-C,” which maps chromatin interactions in nuclei using NGS. The NGS libraries were generated for resting and activated human CD4 T cells to study activation-induced chromatin structural rearrangement. Another chapter (Chapter 13) dealing with chromosomal changes uses CRISPR-based knockout libraries in genome-wide screens to systematically investigate gene function in biological systems. Specifically, the Genome-Scale CRISPR Knock Out (GeCKO v2) library is used and methods for NGS library generation and sequencing are discussed.

There are several chapters (Chapters 7, 8, 10, and 14) dealing with a single cell of DNA or RNA in NGS. Chapter 7 describes a post-bisulfite treatment adapter tagging strategy to generate single-cell data looking at genome-wide cytidine methylation states. Chapter 10 describes the use of modified adapters in a small RNA protocol termed “CleanTag” (TriLink Biotechnologies Inc.) to prepare NGS libraries. The use of these adapters prevents adapter dimer formation, the principal artifactual product in a standard small RNA library prep. The authors demonstrate that the technique can work with very low inputs down to the single-cell level (~10 pg). Chapter 14 uses a single-cell analysis approach to identify and

characterize rare circulating CD4 T cells using a Biomark HD and profiles 96 different genes by quantitative PCR (qPCR) and generates NGS libraries using the Biomark HD Access Array. Chapter 8 describes a protocol for isolation, extraction, and sequencing of single bacterial and archaeal cells using FACS, MDA, a 16S rRNA screen, and a computational approach for quality assurance.

Two other chapters (Chapters 9 and 12) have more of a computational focus. Chapter 9 uses low pass whole genome sequencing and reference-guided assembly of non-model organisms for SNP discovery allowing for the genotyping of populations. Chapter 12 describes a computational pipeline for RNAseq analysis from tissue samples containing a complex heterogeneous population of cell types (e.g., a blood sample). The pipeline is able to estimate cell type composition and other statistical analysis information generated from bulk RNAseq profiles.

Another important component to almost all NGS-related work, especially those involving microbiome studies, is contamination. To this end, Chapter 11 focuses on the best practices and approaches for sample handling, DNA and/or RNA extraction, and library preparation from microbial and viral samples to generate NGS libraries. Chapter 6 discusses a novel method for generating sequencing libraries from viral RNA termed, “Clickseq,” which uses “Click Chemistry” to attach one of the NGS adapters preventing artifactual generation of chimeras, and consequently, greatly increasing the ability to detect rare recombination events in viral RNA. The last chapter in the book (Chapter 16) describes another RNAseq library prep method for profiling reverse transcription termination sites. It is an efficient protocol and generates good NGS library yields from low RNA inputs generated from protocols such as nascent RNA sequencing, RNA Immunoprecipitation (RIPseq), and 5'-RACE structural probing.

Next Generation Sequencing technology has brought together disparate fields of research from bacterial and viral studies to the study of plants, non-model organisms, and of course human disease. This requires collaboration between the users of NGS data and those needed to design and perform the enzymatic procedures for the preparation of sequencing libraries and ensure the desired target is being captured, targeted, or amplified. Other critical players are the engineers that develop microfluidics and sequencing hardware systems or the bioinformatics experts necessary to ensure data generated is being analyzed correctly and translated into a meaningful analysis. We hope you enjoy this book and find it informative and a useful reference.

La Jolla, CA, USA

*Steven R. Head
Phillip Ordoukhanian
Daniel R. Salomon*

Contents

<i>Preface</i>	<i>vii</i>
<i>Contributors</i>	<i>xi</i>
1 An Integrated Polysome Profiling and Ribosome Profiling Method to Investigate In Vivo Translatome	1
<i>Hyun Yong Jin and Changchun Xiao</i>	
2 Measuring Nascent Transcripts by Nascent-seq	19
<i>Fei Xavier Chen, Stacy A. Marshall, Yu Deng, and Sun Tianjiao</i>	
3 Genome-Wide Copy Number Alteration Detection in Preimplantation Genetic Diagnosis	27
<i>Lieselot Deleye, Dieter De Coninck, Dieter Deforce, and Filip Van Nieuwerburgh</i>	
4 Multiplexed Targeted Sequencing for Oxford Nanopore MinION: A Detailed Library Preparation Procedure	43
<i>Timokratis Karamitros and Gkikas Magiorkinis</i>	
5 Hi-Plex for Simple, Accurate, and Cost-Effective Amplicon-based Targeted DNA Sequencing	53
<i>Bernard J. Pope, Fleur Hammet, Tu Nguyen-Dumont, and Daniel J. Park</i>	
6 ClickSeq: Replacing Fragmentation and Enzymatic Ligation with Click-Chemistry to Prevent Sequence Chimeras.	71
<i>Elizabeth Jaworski and Andrew Routh</i>	
7 Genome-Wide Analysis of DNA Methylation in Single Cells Using a Post-bisulfite Adapter Tagging Approach	87
<i>Heather J. Lee and Sébastien A. Smallwood</i>	
8 Sequencing of Genomes from Environmental Single Cells	97
<i>Robert M. Bowers, Janey Lee, and Tanja Woyke</i>	
9 SNP Discovery from Single and Multiplex Genome Assemblies of Non-model Organisms	113
<i>Phillip A. Morin, Andrew D. Foote, Christopher M. Hill, Benoit Simon-Bouhet, Aimee R. Lang, and Marie Louis</i>	
10 CleanTag Adapters Improve Small RNA Next-Generation Sequencing Library Preparation by Reducing Adapter Dimers	145
<i>Sabrina Shore, Jordana M. Henderson, and Anton P. McCaffrey</i>	
11 Sampling, Extraction, and High-Throughput Sequencing Methods for Environmental Microbial and Viral Communities.	163
<i>Pedro J. Torres and Scott T. Kelley</i>	
12 A Bloody Primer: Analysis of RNA-Seq from Tissue Admixtures	175
<i>Casey P. Shannon, Chen Xi Yang, and Scott J. Tebbutt</i>	
13 Next-Generation Sequencing of Genome-Wide CRISPR Screens.	203
<i>Edwin H. Yau and Tariq M. Rana</i>	

14 Gene Profiling and T Cell Receptor Sequencing from Antigen-Specific CD4 T Cells 217
Marie Holt, Anne Costanzo, Louis Gioia, Brian Abe, Andrew I. Su, and Luc Teyton

15 Investigate Global Chromosomal Interaction by Hi-C in Human Naive CD4 T Cells 239
Xiangzhi Meng, Nicole Riley, Ryan Thompson, and Siddhartha Sharma

16 Primer Extension, Capture, and On-Bead cDNA Ligation: An Efficient RNAseq Library Prep Method for Determining Reverse Transcription Termination Sites 253
Phillip Ordoukhanian, Jessica Nichols, and Steven R. Head

Index 263

Contributors

- BRIAN ABE • *Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA, USA*
- ROBERT M. BOWERS • *Department of Energy, Joint Genome Institute, Walnut Creek, CA, USA*
- FEI XAVIER CHEN • *Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA*
- ANNE COSTANZO • *Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA, USA*
- DIETER DE CONINCK • *Laboratory of Pharmaceutical Biotechnology, Ghent University, Ghent, Belgium*
- DIETER DEFORCE • *Laboratory of Pharmaceutical Biotechnology, Ghent University, Ghent, Belgium*
- LIESELOT DELEYE • *Laboratory of Pharmaceutical Biotechnology, Ghent University, Ghent, Belgium*
- YU DENG • *Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA*
- ANDREW D. FOOTE • *Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Bangor University, Bangor, Gwynedd, UK*
- LOUIS GIOIA • *Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA*
- FLEUR HAMMET • *Genomic Technologies Group, Genetic Epidemiology Laboratory, Department of Medicine, The University of Melbourne, Parkville, VIC, Australia*
- STEVEN R. HEAD • *Next Generation Sequencing and Microarray Core Facility, The Scripps Research Institute, La Jolla, CA, USA*
- JORDANA M. HENDERSON • *Research and Development, Cell and Molecular Biology, TriLink BioTechnologies, LLC, San Diego, CA, USA*
- CHRISTOPHER M. HILL • *Department of Computer Science, University of Maryland, College Park, MD, USA*
- MARIE HOLT • *Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA, USA*
- ELIZABETH JAWORSKI • *Department of Biochemistry and Molecular Biology, The University of Texas Medical Branch, Galveston, TX, USA*
- HYUN YONG JIN • *Department of Immunology and Microbial Science, The Scripps Research Institute, La Jolla, CA, USA*
- TIMOKRATIS KARAMITROS • *Department of Zoology, University of Oxford, Oxford, UK; Department of Medical Microbiology, Hellenic Pasteur Institute, Athens, Greece*
- SCOTT T. KELLEY • *Department of Biology, San Diego State University, San Diego, CA, USA*
- AIMEE R. LANG • *Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, La Jolla, CA, USA*

- HEATHER J. LEE • *Epigenetics Programme, Babraham Institute, Babraham Research Campus, Cambridge, UK; School of Biomedical Sciences and Pharmacy, Faculty of Health and Medicine, The University of Newcastle, Callaghan, NSW, Australia*
- JANEY LEE • *Department of Energy, Joint Genome Institute, Walnut Creek, CA, USA*
- MARIE LOUIS • *Scottish Oceans Institute, University of St Andrews, St Andrews, UK*
- GKIKAS MAGIORKINIS • *Department of Zoology, University of Oxford, Oxford, UK*
- STACY A. MARSHALL • *Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA*
- ANTON P. MCCAFFREY • *Research and Development, Cell and Molecular Biology, TriLink BioTechnologies, LLC, San Diego, CA, USA*
- XIANGZHI MENG • *Department of Molecular Medicine, The Scripps Research Institute, La Jolla, CA, USA*
- PHILLIP A. MORIN • *Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, La Jolla, CA, USA*
- TU NGUYEN-DUMONT • *Genomic Technologies Group, Genetic Epidemiology Laboratory, Department of Medicine, The University of Melbourne, Parkville, VIC, Australia*
- JESSICA NICHOLS • *Next Generation Sequencing and Microarray Core Facility, The Scripps Research Institute, La Jolla, CA, USA*
- PHILLIP ORDOUKHANIAN • *Next Generation Sequencing and Microarray Core Facility, The Scripps Research Institute, La Jolla, CA, USA*
- DANIEL JONATHAN PARK • *Melbourne Bioinformatics, The University of Melbourne, Parkville, VIC, Australia; Genomic Technologies Group, Genetic Epidemiology Laboratory, Department of Medicine, The University of Melbourne, Parkville, VIC, Australia*
- BERNARD J. POPE • *Melbourne Bioinformatics, The University of Melbourne, Parkville, VIC, Australia*
- TARIQ M. RANA • *Solid Tumor Therapeutics Program, Moores Cancer Center, University of California, San Diego, La Jolla, CA, USA; Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA; Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA, USA*
- NICOLE RILEY • *Department of Inflammation Biology, La Jolla Institute of Allergy and Immunology, La Jolla, CA, USA*
- ANDREW ROUTH • *Department of Biochemistry and Molecular Biology, The University of Texas Medical Branch, Galveston, TX, USA; Sealy Centre for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, TX, USA*
- CASEY P. SHANNON • *Prevention of Organ Failure (PROOF) Centre of Excellence, Centre for Heart Lung Innovation, University of British Columbia, St. Paul's Hospital, Vancouver, BC, Canada*
- SIDDHARTHA SHARMA • *Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA, USA*
- SABRINA SHORE • *Research and Development, Cell and Molecular Biology, TriLink BioTechnologies, LLC, San Diego, CA, USA*
- BENOIT SIMON-BOUHET • *Centre d'Etudes Biologiques de Chizé, UMR 7372 CNRS-Université de La Rochelle, Villiers-en-Bois, France*
- SÉBASTIEN A. SMALLWOOD • *Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland; Epigenetics Programme, Babraham Institute, Babraham Research Campus, Cambridge, UK*

- ANDREW I. SU • *Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA*
- SCOTT J. TEBBUTT • *Prevention of Organ Failure (PROOF) Centre of Excellence, Centre for Heart Lung Innovation, University of British Columbia, St. Paul's Hospital, Vancouver, BC, Canada; Division of Respiratory Medicine, Department of Medicine, University of British Columbia, Vancouver, BC, Canada*
- LUC TEYTON • *Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA, USA*
- RYAN THOMPSON • *Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA*
- SUN TIANJIAO • *Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA*
- PEDRO J. TORRES • *Department of Biology, San Diego State University, San Diego, CA, USA*
- FILIP VAN NIEUWERBURGH • *Laboratory of Pharmaceutical Biotechnology, Ghent University, Ghent, Belgium*
- TANJA WOYKE • *Department of Energy, Joint Genome Institute, Walnut Creek, CA, USA*
- CHANGCHUN XIAO • *Department of Immunology and Microbial Science, The Scripps Research Institute, La Jolla, CA, USA*
- CHEN XI YANG • *Prevention of Organ Failure (PROOF) Centre of Excellence, Centre for Heart Lung Innovation, University of British Columbia, St. Paul's Hospital, Vancouver, BC, Canada*
- EDWIN H. YAU • *Division of Hematology-Oncology, Department of Internal Medicine, University of California, San Diego, La Jolla, CA, USA; Solid Tumor Therapeutics Program, Moores Cancer Center, University of California, San Diego, La Jolla, CA, USA*

Chapter 1

An Integrated Polysome Profiling and Ribosome Profiling Method to Investigate In Vivo Translatome

Hyun Yong Jin and Changchun Xiao

Abstract

Recent advances in global translatome analysis technologies enable us to understand how translational regulation of gene expression modulates cellular functions. In this chapter, we present an integrated method to measure various aspects of translatome by polysome profiling and ribosome profiling using purified B cells. We standardized our protocols to directly compare the results from these two approaches. Parallel assessment of translatome with these two approaches can generate a comprehensive picture on how translational regulation determines protein output.

Key words Translatome, Ribosome profiling, Ribo-seq, Polysome profiling, Translational regulation of gene expression

1 Introduction

Over the past few decades, quantification of mRNA abundance in polysome fractionation by sucrose density gradient centrifugation (polysome profiling) has been used to investigate translation activity of individual genes. Recently, this was combined with high-throughput technologies, such as microarray or RNA-seq, to reveal genome-wide landscape of translatome [1, 2]. Essentially, global profiling of mRNAs in each fraction of polysome gradient allows the estimation of ribosome occupancy on all mRNAs expressed in a cell. Based on the assumption that ribosome density (the number of ribosomes associated with an mRNA molecule) highly correlates with protein production, the translational efficiency of individual genes can be calculated from these results.

Ribosomes leave ~30 nt footprints (ribosome-protected RNA fragment, RPF) when they are bound to mRNAs [3]. This, in combination with advanced sequencing technology, led to the development of ribosome profiling [4]. In this method, deep sequencing of RPFs would globally quantify the abundance of

ribosome footprints on individual genes and pinpoint the position of ribosomes on mRNAs.

Both technologies can be utilized to investigate global translational activity, but they have distinct strengths over each other [5]. Polysome profiling measures ribosome density of each mRNA, so that translational activity can be directly captured by a single measurement. The readout of ribosome profiling, ribosome footprint abundance, is a sum of mRNA changes and ribosome density changes, so that translational activity is indirectly estimated by subtracting mRNA abundance changes from ribosome footprint abundance changes [5]. On the other hand, ribosome profiling captures positional information of ribosome footprints at the sub-codon level while polysome profiling does not, and is therefore more suitable for investigating alternative start codons or open reading frames [5]. The analysis of RPF distribution also reveals codon-specific regulation of gene expression during the translational elongation stage [6]. The size of ribosome footprint, which can be measured only by ribosome profiling, may indicate differential conformation of ribosomes or the presence of disomes [7, 8]. A comparison of these two technologies is summarized in Table 1.

The most immediate use of these technologies is to measure global gene expression. Historically, mRNA abundance was often used as a proxy for gene expression. However, it has been frequently shown that mRNA abundance and protein level lack sufficient correlation due to the regulation of gene expression at the level of translation [9]. Therefore, faithful measurement of translome is essential for investigating the end point of gene expression, protein abundance, in a cell. Indeed, quantifying translomes by ribosome profiling and polysome profiling led to the discovery of novel genes that are mainly regulated at the translational level in specific cellular contexts [10, 11].

Despite the significance of these technologies, both approaches exhibit technical limitations (Table 1). In polysome profiling, the separation of heavy polysomes starts to lose its accuracy when the number of ribosomes associated with an mRNA exceeds seven or eight. In ribosome profiling, the length of footprints is much smaller (~30 nts) than that of RNA fragments generated by standard RNA-seq methods. This increases the propensity to align the footprints to multiple genomic locations and makes the dataset intrinsically noisy. Moreover, a small fraction of ribosome footprints are aligned to genomic locations outside of coding sequences, such as 5'UTR and 3'UTR, independent of their protein coding capacity [8, 12]. This may result in underestimation of the differential translational regulation occurring in the coding sequences.

To compensate for these limitations, it may be critical to study translome with these two approaches in parallel and cross-validate each other. We recently combined these two approaches to investigate primary B cells with transgenic miR-17~92 expression

Table 1
The strength and weakness of polysome profiling and ribosome profiling

Method	Strength	Weakness
Polysome profiling	<ul style="list-style-type: none"> • Direct measurement of ribosome density per mRNA. • Well-established protocol. 	<ul style="list-style-type: none"> • Low resolution for heavy polysome fractions. • Presence of pseudopolysomes
Ribosome profiling	<ul style="list-style-type: none"> • Retain positional information of ribosome at the sub-codon level. Suitable for revealing alternative translation initiation sites or codon-specific translational regulation. • Retain footprint length information, which can indicate different conformations of ribosome or the presence of disomes. 	<ul style="list-style-type: none"> • Indirect estimation of translation efficiency. Results need to be normalized against mRNA abundance. • Short read length (~30 nts). Ribosome footprints can be aligned to multiple genetic locations. • Footprints present in 5'UTR or 3'UTR may diminish the differential footprint abundance in coding region and underestimate the contribution of translational regulation.

or complete deletion of the miR-17~92 family miRNAs and revealed that the predominant miRNA effect on their target genes occurs at the translational level [13]. We noticed that the current protocols for ribosome profiling and polysome profiling are from different labs and therefore different in terms of sample preparation and detailed procedures. This may hinder the direct comparison between the results from these two approaches. Here, we provide an integrated ribosome profiling and polysome profiling method to directly compare the results from these two approaches (Fig. 1).

2 Materials

2.1 Generation of Sucrose Gradients

1. RNase inhibitor.
2. 100× Cyclohexamide (CHX): 10 mg/mL.
3. 5× sucrose base buffer: 0.4 M NaCl, 25 mM MgCl₂, 100 mM Tris-HCl, pH 7.4, 5 mM DTT, 20 units of RNase inhibitor/mL.
4. Thin wall polyallomer centrifuge tubes, 14 × 89 mm.

2.2 B Cell Purification and Activation

1. MACS LD column (Miltenyi Biotech, San Diego, CA, USA).
2. B cell medium: DMEM-GlutaMAX, 50 mL FCS, 5 mL of 100× Nonessential amino acids, 5 mL of 1 M HEPES, 5 mL of 100× Penn/Strep, 0.6 mL of 55 mM β-Mercaptoethanol.
3. Lipopolysaccharides from *Escherichia coli* 055:B5 (LPS).

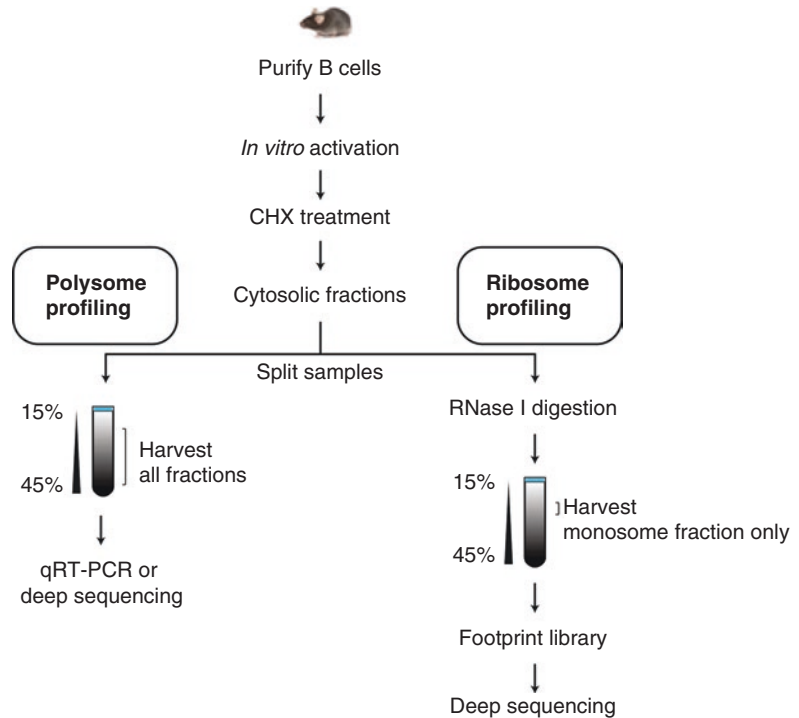


Fig. 1 Concurrent polysome profiling and ribosome profiling to investigate translatoome in vivo

4. IL-4.
5. Ficoll solution.

2.3 Cytosolic Fractionation

1. Hypotonic buffer: 1.5 mM KCl, 10 mM MgCl₂, 5 mM Tris-HCl, pH 7.4, 100 µg/mL CHX.
2. Hypotonic lysis buffer: 2% sodium deoxycholate, 2% Triton X-100, 2.5 mM DTT, 100 µg/mL CHX, 10 units of RNase inhibitor/mL.

2.4 Polysome Profiling and Total RNA Extraction from Each Fraction

1. Solaris control RNA (Dharmacon-GE Healthcare, Lafayette, CO, USA) or ERCC control RNA (ThermoFisher Scientific, Waltham, MA, USA).
2. Trizol-LS (ThermoFisher Scientific).

2.5 Ribosome Profiling Library Preparation

1. RNase decontamination wipes and reagent (*see Note 1*).
2. RNase-free tubes with a low-binding affinity for nucleic acids (*see Note 2*).
3. Hypotonic lysis buffer without RNase Inhibitor: 2% sodium deoxycholate, 2% Triton X-100, 2.5 mM DTT, 100 µg/mL CHX.
4. RNase I.

5. RNase inhibitor.
6. miRNeasy RNA purification kit (Qiagen, Valencia, CA, USA) (*see Note 3*).
7. 2× denaturing loading buffer: 98%(vol/vol) formamide with 10 mM EDTA and 300 µg/mL bromophenol blue.
8. 5× TBE solution: Combine; 27 g Tris Base, 13.7 g Boric acid, pH 8.0, 10 mL 0.5 M EDTA and fill up to 500 mL RNase-free water.
9. 10% polyacrylamide TBE-Urea gel: Combine; 16.8 g urea, 4 mL of 5× TBE and 16 mL RNase-free water. Heat to dissolve urea for 20 min. Cool down the solution to room temperature and add 10 mL of 40% acrylamide/bis solution (37.5:1), 240 µL of 10% APS, and 32 µL of TEMED. For more detailed condition, *see ref. 14*.
10. 10 bp DNA ladder.
11. 26–34 nt RNA marker: 10 µM mix of NI-NI-19 and NI-NI-20. For the detailed sequence information, *see ref. 15*.
12. SYBR gold, 10,000× (ThermoFisher Scientific).
13. Isopropanol (≥99.5%).
14. Glycoblu, 15 mg/mL (ThermoFisher Scientific).

2.5.1 Ribo-Zero rRNA Depletion

1. Ribo-Zero rRNA removal kit (Illumina, San Diego, CA, USA).
2. Zymo RNA Clean and Concentrator-5 kit (Zymo Research, Irvine, CA, USA).

2.5.2 Dephosphorylation, Linker Ligation, and Purification of Linker Ligated Products

1. T4 Polynucleotide Kinase.
2. Linker-1 (/5rApp/CTGTAGGCACCATCAAT/3ddC/) (*see Note 4*).
3. T4 RNA ligase 2, truncated (New England Biolabs, Ipswich, MA, USA), supplied with PEG 8000 50% (wt/vol) and 10× T4 Rnl2 buffer.

2.5.3 Reverse Transcription and Circularization

1. RT primer (*see ref. 15*):
 - (a) 5'-(Phos)-agatcgggaagagcgtcgttagggaaagagtgtagatctcggggtcgc-(iSp18)-cactca-(iSp18)-ttcagacgtgtgctcttccgatc-tattgatgggtcctacag-3'. The designation (Phos) indicates 5'-phosphorylation and -(iSp18)- indicates a 18 atom hexa-ethyleneglycol spacer.
2. 3 M sodium acetate, pH 5.5.
3. 1 N NaOH.
4. 0.1 M Dithiothreitol (DTT).
5. 10 mM dNTP mix.
6. Phusion polymerase (New England Biolabs) (*see Note 5*).

7. 7.5% polyacrylamide TBE-Urea gel: Combine; 16.8 g urea, 4 mL of 5× TBE and 18.5 mL RNase-free water. Heat to dissolve urea for 20 min. Cool down the solution to room temperature and add 7.5 mL of 40% acrylamide/bis solution (37.5:1), 240 μL of 10% APS and 32 μL of TEMED.
8. DNA gel extraction buffer: 300 mM NaCl, 10 mM Tris-HCl, pH 8.0, and 10 mM EDTA.
9. CircLigase (Illumina).

2.5.4 Indexed Library Generation

1. Index library primers (*see* ref. 15):

Barcode_Forward	5'-aatgatacggcgaccaccgagatctacac-3'
Barcode_Reverse_01_ACGACT	5'-caagcagaagacggcatacagatAGTCGTgtgactggagttcagacgtgtgctcttccgatct-3'
Barcode_Reverse_02_ATCAGT	5'-caagcagaagacggcatacagatACTGATgtgactggagttcagacgtgtgctcttccgatct-3'
Barcode_Reverse_03_CAGCAT	5'-caagcagaagacggcatacagatATGCTGgtgactggagttcagacgtgtgctcttccgatct-3'
Barcode_Reverse_04_CGACGT	5'-caagcagaagacggcatacagatACGTCGgtgactggagttcagacgtgtgctcttccgatct-3'
Barcode_Reverse_05_GCAGCT	5'-caagcagaagacggcatacagatAGCTGCgtgactggagttcagacgtgtgctcttccgatct-3'
Barcode_Reverse_06_TACGAT	5'-caagcagaagacggcatacagatATCGTAgtgactggagttcagacgtgtgctcttccgatct-3'
Barcode_Reverse_07_CTGACG	5'-caagcagaagacggcatacagatCGTCAGgtgactggagttcagacgtgtgctcttccgatct-3'
Barcode_Reverse_08_GCTACG	5'-caagcagaagacggcatacagatCGTAGCgtgactggagttcagacgtgtgctcttccgatct-3'
Barcode_Reverse_09_TGCAGC	5'-caagcagaagacggcatacagatGCTGCAGgtgactggagttcagacgtgtgctcttccgatct-3'

2. Phusion polymerase (New England Biolabs).
3. 8% polyacrylamide TBE gel: Combine; 27.8 mL distilled water, 4 mL of 5× TBE, 8 mL of 40% acrylamide:bis solution (37.5:1), 200 μL of APS and 20 μL of TEMED.
4. 6× DNA gel loading dye.

2.6 Illumina Sequencing

1. Standard Illumina cluster generation and sequencing kits (minimum 50 bases plus indexing reads) appropriate for the sequencer used (HiSeq, MiSeq, or NextSeq).

3 Methods

3.1 Generation of Sucrose Gradients

1. Generate 15% sucrose solution (50 mL of 5× sucrose base buffer, 37.5 g of sucrose, and fill up to 250 mL).
2. Generate 45% sucrose solution (50 mL of 5× sucrose base buffer, 112.5 g of sucrose, and fill up to 250 mL).
3. (Optional) Filter the sucrose solution using 0.45 µm vacuum filter.
4. Next, generate five different concentrations of gradients by mixing 15% and 45% sucrose solution.

	15% sucrose solution	45% sucrose solution
Final 15% sucrose solution	50 mL	0 mL
Final 22.5% sucrose solution	37.5 mL	12.5 mL
Final 30% sucrose solution	25 mL	25 mL
Final 37.5% sucrose solution	12.5 mL	37.5 mL
Final 45% sucrose solution	0 mL	50 mL

5. Add 2.2 mL of 45% sucrose gradient to thin wall polyallomer centrifuge tubes and freeze it down for 20 min at -80°C . Once it is frozen, overlay with 2.2 mL of 37.5% sucrose gradient and freeze it down for 20 min. Repeat it with 30%, 22.5% and 15% sucrose gradient. This will generate frozen 15–45% gradient stocks. They can be thawed in a cold room (4°C) for a few hours to up to 12 h before centrifugation is conducted.

3.2 B Cell Purification and Activation

1. 100×10^6 follicular B cells were purified from total splenocytes by depleting cells positive for CD5, CD43 or CD93 using MACS LD column according to the manufacturer's instruction. The purified cells were in vitro activated with LPS (25 µg/mL) and IL-4 (5 ng/mL) in B cell medium for aimed time points [16].
2. Before harvest, add CHX directly to the culture medium and incubate for 15 min. This will freeze the ribosomes on mRNAs.
3. (Optional) At later time points of B cell activation, a significant fraction of cells undergo apoptosis. Purifying live cells from the culture may improve the results. Live cells can be purified using Ficoll solution according to the manufacturer's instruction. During this process, add same final concentration of CHX to Ficoll solution to maintain ribosome association to mRNAs.
4. Split samples into two groups at this point. One for polysome profiling (*see* Subheading 3.3) and the other for ribosome profiling (*see* Subheading 3.5.1).

3.3 Cytosolic Fractionation

1. Wash cells with 10 mL of ice-cold hypotonic buffer.
2. Swell cells for 10 min on ice.
3. Spin down with $500 \times g$ for 5 min at 4 °C, aspirate the hypotonic buffer, and gently resuspend the cells using 300 μ L hypotonic buffer.
4. Transfer the resuspended cells to new tube and add 300 μ L hypotonic lysis buffer (total 600 μ L).
5. Incubate samples on ice for 20 min.
6. Spin down the cells with $2300 \times g$ at 4 °C for 10 min and collect supernatant. This removes nuclei and collects the cytosolic fraction.
7. (*Optional*) Confirm the cytosolic fractionation efficiency (*see* Fig. 2).

3.4 Polysome Profiling and Total RNA Extraction from Each Fraction

1. Carefully overlay the cytosolic fractions on the thawed 15–45% sucrose gradient at 4 °C.
2. Centrifuge at 40,000 rpm ($274,000 \times g$ at r-max and $121,000 \times g$ at r-min) for 1.5 h at 4 °C, using SW41 rotor.
3. Fractionate the sucrose into 20 tubes using automated fractionation machine and simultaneously measure A254 values to estimate the overall translation profile. Each fraction will contain roughly 540 μ L of sucrose after fractionation. Flow rates of 0.75 mL/min and collection exchange speed of 0.73 min/tube are recommended.
4. (*Optional*) Add equal amounts of Solaris RNA (2 μ L) to each collection as a normalization control. Alternatively, ERCC RNA can be used.
5. Add 1.5 mL of Trizol-LS to each tube (total ~2 mL solution) and extract total RNA following the manufacturer's instruction.

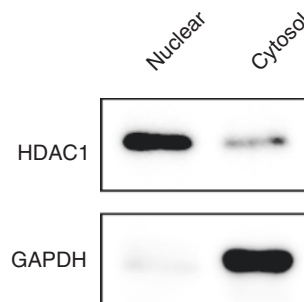


Fig. 2 Western blot analysis showing the efficient separation of cytosolic from nuclear fractions. After the cytosolic fractions are harvested, remaining pellets (nuclear fractions) were further lysed with 1% NP-40 lysis buffer with standard procedure. Note that the nuclear marker, HDAC1, is mainly present in the nucleus while the cytosolic marker, GAPDH, is present exclusively in cytosolic fractions

6. Conduct standard qRT-PCR or RNA-seq of each fraction to reveal polysome distribution of individual genes or global polysome distribution, respectively.

3.5 Ribosome Profiling Library Preparation

Ribosome footprint library is generated as previously described with modifications to make the conditions as comparable to that of polysome profiling as possible [4, 15]. Major modifications are: (1) Used total RNA purified from cytosolic fractions instead of whole cell lysate with DNase treatment method [15]. (2) RNase I treated footprints are purified using the same sucrose gradient as polysome profiling. This method was originally proposed in the initial ribosome profiling protocol [4], but later modified to either sucrose cushion method [15] or Sephacryl S-400 column purification method [17, 18]. (3) rRNAs are depleted using Ribo-zero rRNA depletion kit instead of biotinylated rRNA depletion oligos after footprint recovery and before the dephosphorylation and linker ligation steps [15].

3.5.1 Cytosolic Fractionation

1. Extract the cytosolic fraction (final volume of 600 μ L) as described in Subheading 3.3, using the hypotonic lysis buffer without RNase Inhibitor. Maintain 4 °C in cold room to avoid unwanted mRNA degradation during the process.

3.5.2 RNase I Footprinting and Ribosome Recovery

1. Take the 600 μ L of cytosolic fraction and add 7.5 μ L of RNase I (100 U/ μ L).
2. Incubate for 45 min at room temperature with gentle mixing on nutator.
3. To stop the RNase I digestion, add 10 μ L of SUPERase Inhibitor and tap the tubes gently.
4. Run 15–45% sucrose gradients as described in Subheading 3.4 and confirm efficient RNase I digestion (*see* Fig. 3).
5. Collect and combine fractions 6, 7, and 8. Fractions 9–10 include disome-protected mRNA fragments which may represent ribosome stalling [8], and are excluded from our analysis (*see* Fig. 3).
6. Add 4.5 mL of Trizol-LS to the combined fractions (total 6 mL) and extract total RNA. To increase yield, the miRNeasy RNA purification kit is recommended. Expected amounts of RNA from 50 million B cells in each step are listed in Table 2.

3.5.3 Footprint Gel Purification

1. Add 2 \times denaturing loading buffer to each sample.
2. Prepare 1 μ L of 26 nt and 34 nt RNA marker and 1 μ L of 10 bp DNA ladder as control. Add 4 μ L of 10 mM Tris-HCl (pH 8.0) and 5 μ L of 2 \times denaturing loading buffer to each.
3. Denature the sample at 80 °C for 90 s and transfer the sample to ice directly.

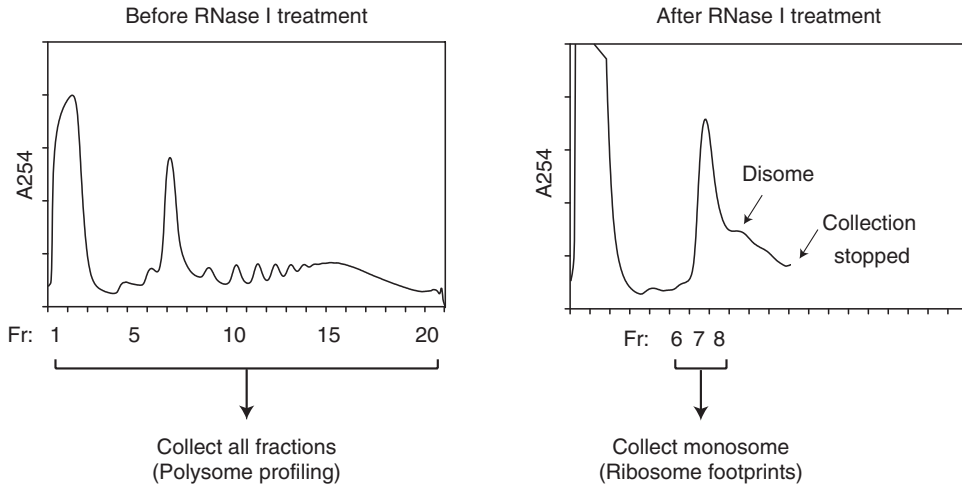


Fig. 3 A254 profiles before and after RNase I digestion. Profiles are from 25.5h activated wild type B cells

Table 2
Expected RNA amount at each step, starting with 50 million 25.5 h activated B cells

Step	Method	Expected RNA amount
3.5.1	Total RNA from cytosolic fractions and miRNeasy purification	~200 µg
3.5.2	RNase I treated, selection of ribosome protected monosome from sucrose gradient and miRNeasy purification	~15 µg
3.5.3	26–32 nt RNA gel purification	~1 µg
3.5.4	rRNA depletion using Ribo-zero kit	~200 ng
3.5.5	Linker ligation	90% efficiency

4. Separate the sample using 10% polyacrylamide TBE-Urea gel in 0.5× TBE buffer with constant power (5 W) for 80 min until the bromophenol blue dye reaches two third of the gel. Flushing each well with 0.5× TBE solution with syringe before sample loading helps to generate nicely shaped bands.
5. Carefully disassemble the gel cassette and stain the gel for 3 min with 1× SYBR gold in 0.5× TBE buffer (*see Fig. 4*).
6. Excise the ribosome footprint (24–36 nt region) and 26 nt/34 nt marker (*see Fig. 4*). 26 nt/34 nt footprints extracted here serve as a control for the remaining steps.
7. Overnight gel extraction. Add 400 µL of RNA gel extraction buffer and freeze the sample for 30 min on dry ice. Leave the

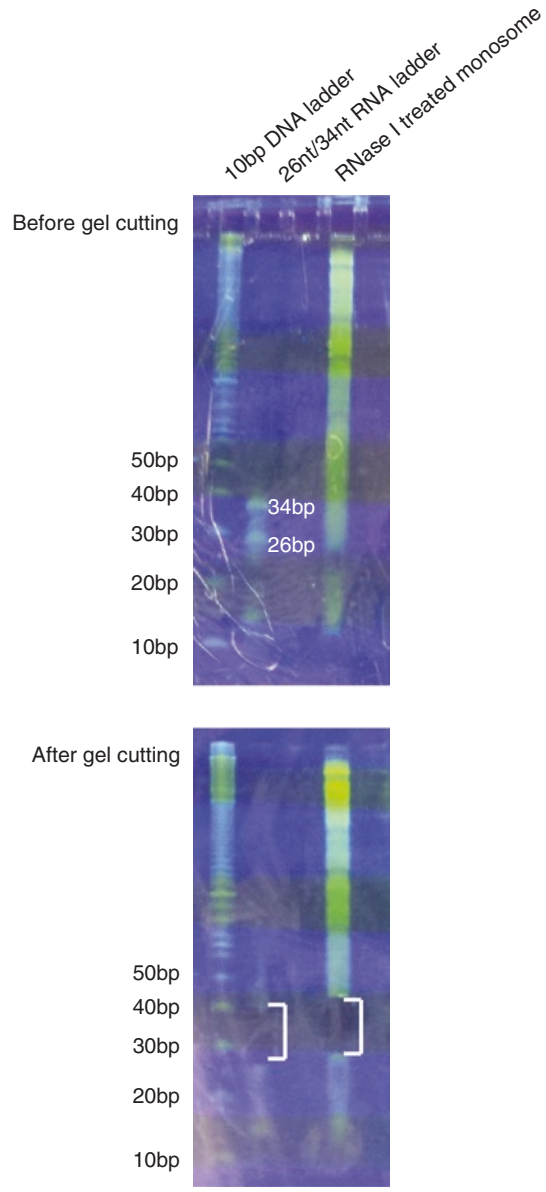


Fig. 4 Size selection of ribosome footprints from RNase I treated monosome fractions. Representative gel images before and after gel cutting are shown. Brackets indicate the location of excised gel bands

sample overnight (15 h is recommended) with gentle mixing. Collect all solution and transfer to non-stick RNase-free eppitubes. Precipitate the footprints by adding 2 μL Glycoblue and 500 μL isopropanol. Carry out precipitation for 30 min or more on dry ice. Centrifuge at maximum speed at 4 $^{\circ}\text{C}$ for 30 min using a tabletop centrifuge. Footprints frequently end up on the side of the tubes. If this happens, resuspend footprints and centrifuge again. Completely remove the supernatant and allow air dry for 5 min. Dissolve footprints in 10 μL RNase-free water.

3.5.4 Ribo-Zero rRNA Depletion

1. Conduct the rRNA depletion using Ribo-zero rRNA removal kit following the manufacturer's protocol.
2. Concentrate rRNA-depleted sample using Zymo RNA Clean and Concentrator kit according to the manufacturer's protocol. At the final elution step, use 8 μL of RNase-free water, which will give 7 μL final elute.

3.5.5 Dephosphorylation, Linker Ligation, and Purification of Linker Ligated Products

1. Dephosphorylation reaction. Mix the 7 μL of rRNA depleted footprint and 1 μL PNK buffer. Incubate at 75 $^{\circ}\text{C}$ for 1 min and cool on ice. Add 1 μL of SUPERase Inhibitor and 1 μL T4 PNK. Incubate for 1 h at 37 $^{\circ}\text{C}$ (total volume: 10 μL). Upon completion, heat-inactivate samples at 65 $^{\circ}\text{C}$ for 3 min and cool samples on ice.
2. Linker ligation reaction. Mix the 10 μL of dephosphorylated products and 1 μL of linker-1. Denature the sample for 90 s at 80 $^{\circ}\text{C}$ and immediately cool on ice for 10 min. Add 1 μL of T4 Rnl2 buffer, 1 μL of SUPERase inhibitor, 6 μL of PEG8000, and 1 μL of T4 RNA ligase 2 (truncated). Incubate for 3 h at 25 $^{\circ}\text{C}$.
3. Add 30 μL of RNase-free water to each sample and purify ligated product using Zymo RNA Clean and Concentrator-5 kit. Note that the kit efficiently removes un-ligated linker (*see* Fig. 5). At the final step, elute the ligated product in 11 μL of 10 mM Tris-HCl (pH 8.0) solution. Final elute is \sim 10 μL .

3.5.6 Reverse Transcription and Circularization

1. Add 2 μL of 0.25 μM reverse transcription primer to 10 μL of linker-ligated product.
2. Denature the sample for 2 min at 80 $^{\circ}\text{C}$ and immediately cool on ice.
3. Add 8 μL of reverse transcription master mix to each sample to make a 20 μL reaction (Master mix: 4 μL of first-strand buffer, 1 μL of 10 nM dNTP, 1 μL of 0.1 M DTT, 1 μL of SUPERase inhibitor, and 1 μL of Super Script III).
4. Hydrolyze the RNA template by adding 2.2 μL of 1 N NaOH and incubate at 90 $^{\circ}\text{C}$ for 20 min.

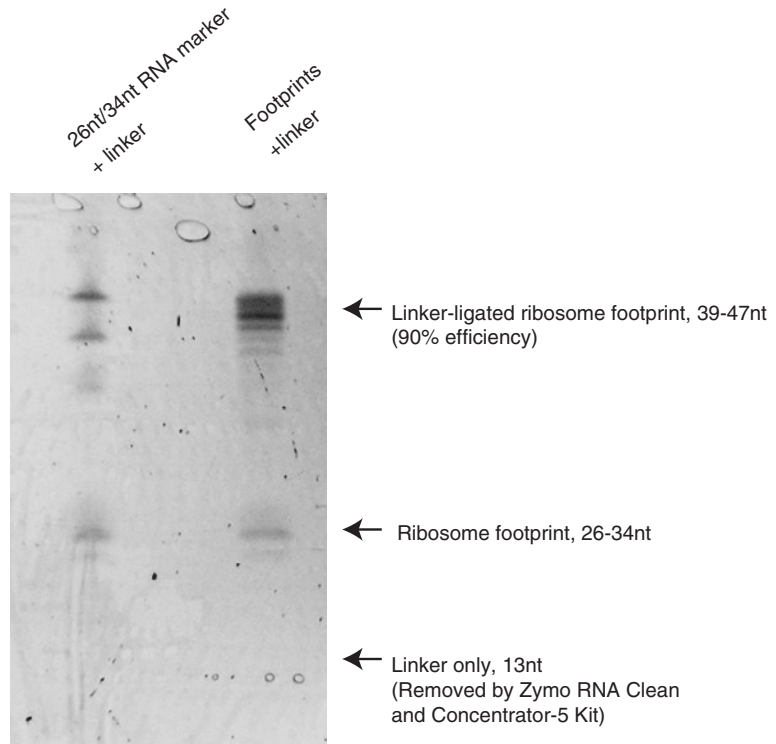


Fig. 5 Confirm the efficiency of linker ligation. Zymo RNA Clean and Concentrator-5 purified linker ligated RNA marker or footprints. Un-ligated linkers are completely removed by the Zymo kit, therefore do not need this gel running steps for experimental samples

5. Precipitate the RT product. Add 20 μL of 3 M sodium acetate (pH 5.5), 2 μL of Glycoblue, 156 μL distilled water, 300 μL isopropanol and precipitate on dry ice for 30 min. Pellet the DNA for 30 min as described and resuspend the RT product in 10 μL distilled water.
6. Prepare 2 μL of 0.25 μM RT primer (mixed with 3 μL of 10 nM Tris-HCl, pH 8.0 and 5 μL of 2 \times denaturing loading buffer) and 1 μL of 10 bp DNA ladder as control. Add 10 μL of 2 \times loading buffer to each sample. Denature the sample at 80 $^{\circ}\text{C}$ for 90 s and immediately cool on ice.
7. Separate the sample using 7.5% polyacrylamide TBE-Urea gel in 0.5 \times TBE buffer with constant power (5 W) for 1 h and 30 min until the bromophenol blue dye reaches the end of the gel (*see* Fig. 6).
8. Excise the reverse transcription product and transfer to non-sticky eppi-tubes. 2 mL tubes are recommended for DNA extraction. Avoid any contamination from un-extended RT primer.

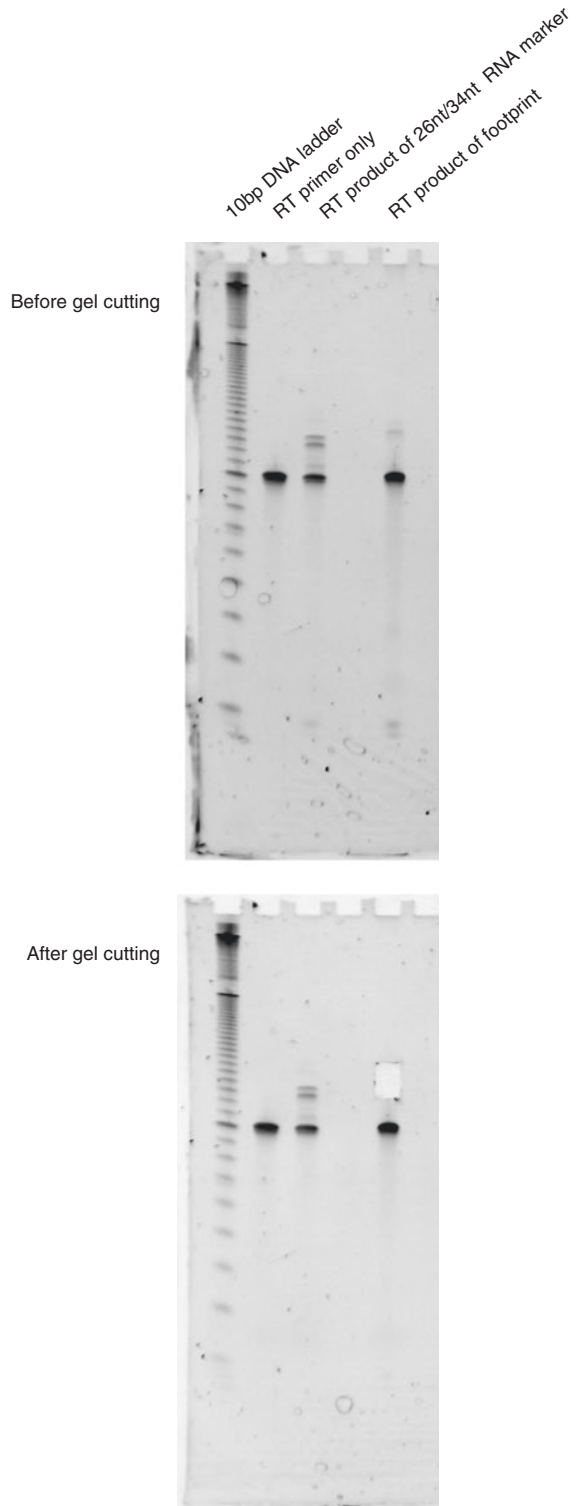


Fig. 6 Representative gel picture showing RT product of linker-ligated 26nt/34nt RNA marker and footprint

9. Overnight gel extraction. Add 400 μL of DNA gel extraction buffer and freeze the sample for 30 min on dry ice. Leave the sample overnight (15 h is recommended) with gentle mixing. Collect the entire solution and transfer it to non-stick RNase-free tubes (*see Note 2*). Precipitate the footprints by adding 2 μL Glycoblue and 500 μL isopropanol. Carry out precipitation for 30 min or more on dry ice. Centrifuge at maximum speed at 4 $^{\circ}\text{C}$ for 30 min using tabletop centrifuge. Completely remove the supernatant and allow air dry for 5 min. Dissolve the footprint in 15 μL of 10 mM Tris-HCl (pH 8.0).
10. Circularization reaction according to the manufacturer's protocol. Mix 15 μL of gel extracted product, 2 μL of CircLigase buffer, 1 μL of 1 mM ATP, 1 μL of 10 nM MnCl_2 and 1 μL of CircLigase per sample, incubate the reaction for 1 h at 60 $^{\circ}\text{C}$ and heat-inactivate for 10 min at 80 $^{\circ}\text{C}$.
11. Recover the final DNA product. Add 74 μL distilled water, 6 μL of 5 M NaCl, 2 μL of Glycoblue and 150 μL of isopropanol and precipitate DNA as described in **step 9** in this session.
12. Resuspend the circularized DNA product in 5 μL of 10 mM Tris-HCl (pH 8.0).

3.5.7 Indexed Library Generation

1. PCR amplification of circularized library using Phusion polymerase according to the manufacturer's instruction. Run several different cycles (i.e., 4, 6, and 8 cycles) for comparison. Use different barcode reverse primers (Barcode_Reverse_##) for different samples with a single forward primer (Barcode_Forward).
2. Add 6 \times DNA loading dye to each PCR product and prepare 1 μL 10 bp DNA ladder as control.
3. Separate the PCR products using 8% polyacrylamide TBE gel, for 90 min at 180 V in 0.5 \times TBE buffer, or until the bromophenol blue dye reaches to two-third of the gel (*see Fig. 7*).
4. Stain the gel for 3 min in 1 \times SYBR Gold in 0.5 \times TBE buffer.
5. Excise the \sim 175 nt PCR product from the gel and carefully exclude slowly migrating smeared band (above 200 nt), bands from un-extended reverse transcription (\sim 145 nt), template and primer (less than 100 nt) (*see Fig. 7*).
6. Overnight gel extraction. Add 400 μL of DNA gel extraction buffer and freeze the sample for 30 min on dry ice. Leave the sample overnight (15 h is recommended) with gentle mixing. Collect the entire solution and transfer it to non-stick RNase-free eppi tubes. Precipitate the footprints by adding 2 μL Glycoblue and 500 μL isopropanol. Carry out precipitation for 30 min or more on dry ice. Centrifuge at maximum speed at 4 $^{\circ}\text{C}$ for 30 min using tabletop centrifuge. Completely remove

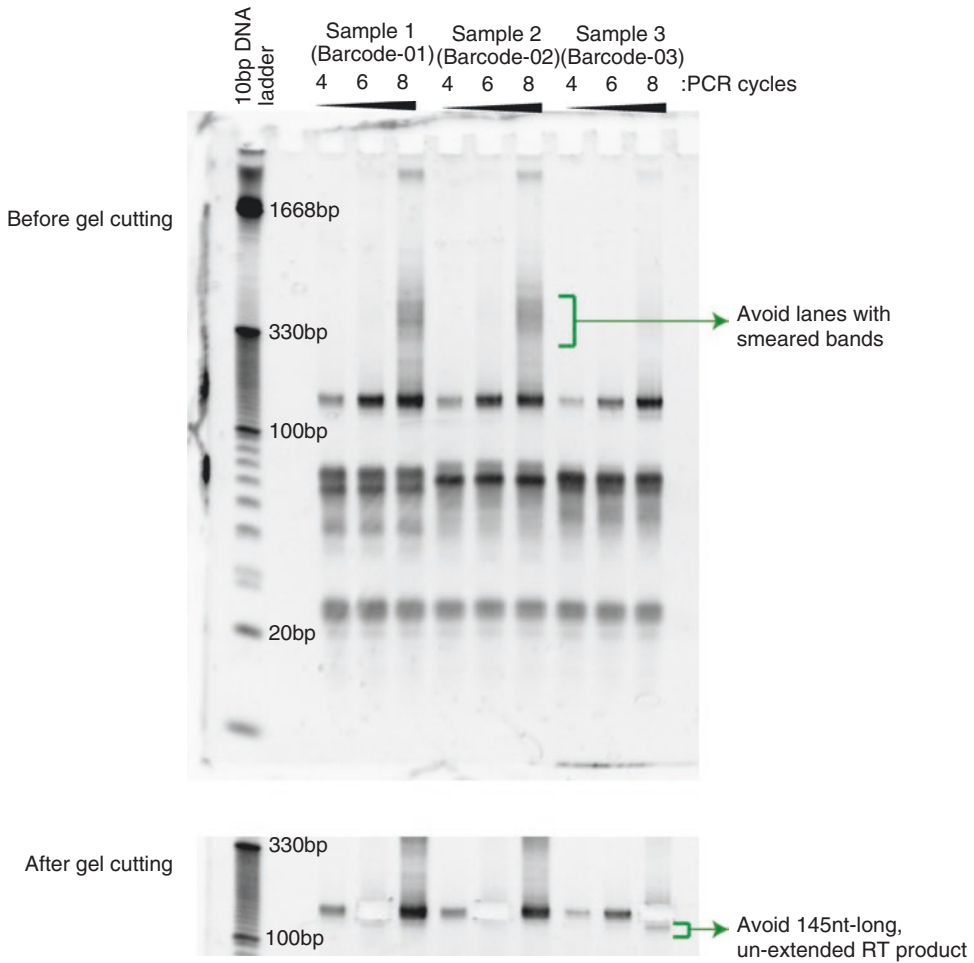


Fig. 7 Representative gel pictures of barcoded PCR products

the supernatant and air dry for 5 min. Dissolve the footprint in 15 μ L of 10 mM Tris-HCl (pH 8.0).

7. Measure DNA concentration and combine equal amount from each sample for Illumina sequencing.

3.6 Illumina Sequencing

1. The end product of this protocol results in a fully functional sequencing library compatible with Illumina HiSeq, MiSeq, and NextSeq sequencing platforms. The user should follow standard procedures recommended by Illumina for adjusting the final library concentration for loading onto an Illumina flow-cell and sequencing appropriate for the sequencer used. We recommend a minimum of single-end 50 bp read lengths with six base indexing reads for downstream demultiplexing as described here. Sequencing protocols for polysome fraction and ribosome profiling are identical.

4 Notes

1. Maintain RNase-free working conditions: Routine use of RNase decontamination wipes, e.g., RNaseZAP or RNase Displace (ThermoFisher Scientific, Waltham, MA, USA) to wipe down the work area and rinse gel apparatuses is recommended.
2. Non-stick, RNase-free tubes are recommended throughout protocol, e.g., Eppendorf[®] DNA LoBind microcentrifuge tubes (Eppendorf, Hamburg, Germany) or Nonstick, RNase-free Microfuge Tubes (ThermoFisher Scientific, Waltham, MA, USA).
3. Other RNA purification columns may be substituted at this step.
4. There is a pre-made version available from Integrated DNA Technologies (Coralville, IA, USA) or a custom synthesis can be done.
5. Other commercially available thermostable high-fidelity polymerases can be substituted for Phusion (e.g., KAPPA Hi-Fi, Herculase II Fusion).

Acknowledgments

We thank Jovan Shepherd for critical reading of manuscript. C.X. is a Pew Scholar in Biomedical Sciences. This study is supported by the PEW Charitable Trusts, Cancer Research Institute, National Institute of Health (R01AI087634, R01AI089854, RC1CA146299, R56AI110403, and R01AI121155 to C.X.).

References

1. Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D (2003) Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* 100(7):3889–3894. <https://doi.org/10.1073/pnas.0635171100>. [pii] 0635171100
2. Lackner DH, Beilharz TH, Marguerat S, Mata J, Watt S, Schubert F, Preiss T, Bahler J (2007) A network of multiple regulatory layers shapes gene expression in fission yeast. *Mol Cell* 26(1):145–155. <https://doi.org/10.1016/j.molcel.2007.03.002>. [pii] S1097-2765(07)00147-5
3. Steitz JA (1969) Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. *Nature* 224(5223):957–964
4. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324(5924):218–223. <https://doi.org/10.1126/Science.1168978>
5. Ingolia NT (2014) Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet* 15(3):205–213. <https://doi.org/10.1038/nrg3645>. [pii] nrg3645
6. Nedialkova DD, Leidel SA (2015) Optimization of codon translation rates via tRNA modifications maintains proteome integrity. *Cell* 161(7):1606–1618. <https://doi.org/10.1016/j.cell.2015.05.022>. [pii] S0092-8674(15)00571-1
7. Lareau LF, Hite DH, Hogan GJ, Brown PO (2014) Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *elife* 3:e01257. <https://doi.org/10.7554/eLife.01257>

8. Guydosh NR, Green R (2014) Dom34 rescues ribosomes in 3' untranslated regions. *Cell* 156(5):950–962. <https://doi.org/10.1016/j.cell.2014.02.006>. [pii] S0092-8674(14)00162-7
9. Payne SH (2015) The utility of protein and mRNA correlation. *Trends Biochem Sci* 40(1):1–3. <https://doi.org/10.1016/j.tibs.2014.10.010>. [pii] S0968-0004(14)00202-3
10. Schott J, Reitter S, Philipp J, Haneke K, Schafer H, Stoecklin G (2014) Translational regulation of specific mRNAs controls feedback inhibition and survival during macrophage activation. *PLoS Genet* 10(6):e1004368. <https://doi.org/10.1371/Journal.Pgen.1004368>. Art n E1004368
11. Schafer S, Adami E, Heinig M, Rodrigues KE, Kreuchwig F, Silhavy J, van Heesch S, Simaite D, Rajewsky N, Cuppen E, Pravenec M, Vingron M, Cook SA, Hubner N (2015) Translational regulation shapes the molecular landscape of complex disease phenotypes. *Nat Commun* 6:7200. <https://doi.org/10.1038/ncomms8200>. [pii] ncomms8200
12. Liu BT, Han Y, Qian SB (2013) Cotranslational response to proteotoxic stress by elongation pausing of ribosomes. *Mol Cell* 49(3):453–463. <https://doi.org/10.1016/j.molcel.2012.12.001>
13. Jin HY, Oda H, Chen P, Yang C, Zhou X, Kang SG, Valentine E, Kefauver JM, Liao L, Zhang Y, Gonzalez-Martin A, Shepherd J, Morgan GJ, Mondala TS, Head SR, Kim P-H, Xiao N, Fu G, Liu W-H, Han J, Williamson JR, Xiao C (2017) Differential Sensitivity of Target Genes to Translational Repression by miR-17~92. *PLoS Genet* 13 (2):e1006623. doi: [10.1371/journal.pgen.1006623](https://doi.org/10.1371/journal.pgen.1006623)
14. Jin HY, Gonzalez-Martin A, Miletic A, Lai M, Knight S, Sabouri-Ghomi M, Head SR, Macauley MS, Rickert R, Xiao C (2015) Transfection of microRNA mimics should be used with caution. *Front Genet* 6:340. <https://doi.org/10.3389/fgene.2015.00340>
15. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* 7(8):1534–1550. <https://doi.org/10.1038/nprot.2012.086>. [pii] nprot.2012.086
16. Jin HY, Oda H, Lai M, Skalsky RL, Bethel K, Shepherd J, Kang SG, Liu WH, Sabouri-Ghomi M, Cullen BR, Rajewsky K, Xiao C (2013) MicroRNA-17~92 plays a causative role in lymphomagenesis by coordinating multiple oncogenic pathways. *EMBO J* 32(17):2377–2391. <https://doi.org/10.1038/emboj.2013.178>. [pii] emboj2013178
17. Cho J, Yu NK, Choi JH, Sim SE, Kang SJ, Kwak C, Lee SW, Kim JI, Choi DI, Kim VN, Kaang BK (2015) Multiple repressive mechanisms in the hippocampus during memory formation. *Science* 350(6256):82–87. <https://doi.org/10.1126/science.aac7368>. [pii] 350/6256/82
18. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, Wills MR, Weissman JS (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* 8(5):1365–1379. <https://doi.org/10.1016/j.celrep.2014.07.045>. [pii] S2211-1247(14)00629-9

Chapter 2

Measuring Nascent Transcripts by Nascent-seq

Fei Xavier Chen, Stacy A. Marshall, Yu Deng, and Sun Tianjiao

Abstract

A complete understanding of transcription and co-transcriptional RNA processing events by polymerase requires precise and robust approaches to visualize polymerase progress and quantify nascent transcripts on a genome-wide scale. Here, we present a transcriptome-wide method to measure the level of nascent transcribing RNA in a fast and unbiased manner.

Key words Nascent-seq, Transcription, Pol II, mRNA, Next-generation sequencing

1 Introduction

Transcription is a remarkably dynamic process during which the information in DNA is copied into RNA. High-throughput sequencing has revealed different categories of noncoding RNAs transcribed in coordination with or independent of protein-coding messenger RNA (mRNA) [1, 2]. Transcription of these coding and noncoding RNA molecules requires a precise and efficient collaboration of numerous transcription factors to regulate polymerase recruitment, transcriptional initiation, pause release, elongation, and termination [3, 4]. For transcription of mRNA and a subset of noncoding RNAs, transcription-coupled RNA processing events (e.g., 5' capping, co-transcriptional splicing, and polyadenylation) add an additional layer of complexity to the regulation of transcription [5]. Thus, approaches to precisely measuring the level of nascent transcripts on genome-wide scale are essential for a full understanding of the dynamic process of transcriptional control.

With the help of high-throughput sequencing, various strategies have been developed to measure the global transcriptional activity. In principle, they can be grouped into two categories. The first, Global Run-On sequencing (GRO-seq) [6] and Precision nuclear Run-On sequencing (PRO-seq) [7], is built on the measurement of nascent RNA extended by run-on assay with modified nucleotides *in vitro*. However, the treatment with sarcosyl and resumption of transcription

elongation with labeled nucleotides *in vitro* may introduce biases and misrepresent the native transcription profiles. The second, Native Elongating Transcript sequencing (NET-seq) [8–10] and nascent RNA sequencing (nascent-seq) [11–14], relies on the extraordinary stability of the DNA–RNA–RNA polymerase ternary complex [15]. In NET-seq, nascent transcripts are enriched and purified by immunoprecipitation of actively transcribing endogenous or tagged RNA polymerase on chromatin. For immunoprecipitation of endogenous polymerase, antibodies that recognize polymerases in different states (e.g., Pol II with different phosphorylation at CTD) with the same or similar efficiency are essential to depict the global transcription profiles in a minimally biased manner. Nascent-seq, which we describe in this chapter, was developed on the basis of the fact that engaged polymerase containing nascent RNA tightly associates with chromatin through a stringent wash with high salt and urea buffer [16]. Nascent RNA is then purified from this washed chromatin. Compared with other approaches measuring nascent transcripts, nascent-seq is much simpler in terms of nascent RNA isolation and library preparation, which enables a fast and efficient measurement of nascent transcripts.

The process of nascent-seq roughly contains four procedures: isolation of cell nuclei, purification of nascent RNA, library preparation, and next-generation sequencing. Nuclei isolation and nascent RNA purification can be finished within 1 day. Library preparation is usually split into 2 days (the first day for cDNA preparation and the second day for library preparation). Sequencing takes no more than ~1–2 h to set up and runs for ~11–13 h for reads of 50 bp in length.

2 Materials

2.1 Equipment and Supplies

1. Allegra X-14R Centrifuge (Beckman Coulter, Carlsbad, CA, USA).
2. Microfuge 20R Centrifuge (Beckman Coulter).
3. Wheaton Dounce tissue grinder, 7 mL.
4. UV-Vis Spectrophotometer, e.g., Nanodrop 2000 (ThermoFisher Scientific, Waltham, MA, USA).
5. Nutator.
6. 1.7 mL Microtubes.
7. 8-Tube Strips with Attached Domed Caps, 0.2 mL PCR tubes.
8. Magnetic Rack.
9. Thermal Cycler.
10. 2100 Bioanalyzer or TapeStation (Agilent Technologies, Santa Clara, CA, USA).
11. Qubit® 2.0 Fluorometer, Assay Tubes, and dsDNA HS Assay Kit (ThermoFisher Scientific).

2.2 Reagents and Buffers

1. Phosphate-buffered saline (PBS).
2. Hypotonic Buffer: 10 mM HEPES, pH = 7.9; 10 mM KCl; 2 mM MgCl₂; 1 mM DTT, add before use; 1× protease inhibitor (Sigma, St. Louis, MO, USA), add before use.
3. Nuclei Wash Buffer: 10 mM HEPES, pH = 7.9; 250 mM Sucrose; 1 mM DTT, add before use; 1× protease inhibitor, add before use; 50 U/mL RNase inhibitor (ThermoFisher Scientific).
4. 2× NUN Buffer (prepared in RNase-free water): 40 mM HEPES, pH = 7.9; 15 mM MgCl₂; 0.4 mM EDTA; 600 mM NaCl; 2% v/v Nonidet P40.
5. 1× NUN Buffer: 50% volume of 2 M fresh urea solution; 50% volume of ice-cold 2× NUN Buffer; 50 U/mL RNase inhibitor.
6. Trypan Blue Solution, 0.4% (ThermoFisher Scientific).
7. TRIzol Reagent (ThermoFisher Scientific).
8. Chloroform.
9. Isopropyl alcohol.
10. 75% ethanol (in RNase-free water).
11. DNase I (RNase-free) (New England Biolabs, Ipswich, MA, USA).
12. High Sensitivity DNA Kit (Agilent Technologies).
13. RNA 6000 Nano Kit (Agilent Technologies).
14. RNase-free water.
15. 10 mM Tris-HCl, pH 8.0.
16. 200 mM Tris-HCl, pH 7.0.
17. 1 M Sodium hydroxide solution.
18. AMPure XP-PCR Purification (Beckman Coulter).
19. RNAClean XP with Scalable throughput (Beckman Coulter).
20. SuperScript II Reverse Transcriptase (ThermoFisher Scientific).
21. TruSeq® Stranded Total RNA LT—(with Ribo-Zero™ Human/Mouse/Rat) (Illumina, San Diego, CA, USA).

3 Methods

3.1 Preparation of Cell Cultures

1. Culture 1–5 × 10⁷ cells for each nascent RNA-seq sample (*see Note 1*).
2. For adherent cells, harvest the cells with trypsin-EDTA or by scraping in PBS; for suspension cells, pellet the cells by spinning at 300 × *g* for 5 min. Discard the supernatant.
3. Resuspend the pellet with 10 mL of ice-cold PBS and transfer the cells to a 15 mL conical tube. Spin at 300 × *g* for 5 min at

4 °C to pellet cells. Repeat this wash two more times. Cell pellet can be used directly for nuclei isolation or be flash-frozen in liquid nitrogen and stored at -80 °C until use, with the former recommended to avoid the interference in nuclei isolation by the freezing and thawing of cells.

3.2 Isolation of Cell Nuclei

1. Resuspend the cell pellet in 10 mL of ice-cold Hypotonic Buffer with protease inhibitor and DTT. Incubate the cell suspension on ice for 15 min.
2. Centrifuge the cell suspension at $200 \times g$ for 10 min at 4 °C. Discard the supernatant.
3. Resuspend the pellet in 5 mL of cold Hypotonic Buffer with protease inhibitor and DTT.
4. Transfer cell suspension into a precooled 7 mL Dounce tissue grinder. Homogenize the suspension with tight pestle with ~10 strokes. Check cell disruption under the microscope with Trypan Blue staining. Aim for ~90% disruption (Disrupted cells can take up the dye, unbroken cells cannot). Do not over-disrupt the cells.
5. Pellet the nuclei by spinning at $600 \times g$ for 10 min at 4 °C. Discard the supernatant that contains the cytoplasmic fraction.
6. Resuspend the pellet with 5 mL of ice-cold Nuclei Wash Buffer with DTT, protease and RNase inhibitor. Spin at $1500 \times g$ for 5 min at 4 °C and discard the supernatant. Repeat this wash once more.

3.3 Purification of Nascent RNA

1. Prepare 1× NUN buffer by mixing 50% volume of 2 M urea solution (fresh and filtered) and 50% volume of ice-cold 2× NUN Buffer supplemented with DTT, protease and RNase inhibitor (*see Notes 2 and 3*).
2. Suspend the pellet and disrupt the nuclei with 1 mL of 1× NUN buffer by pipetting up and down vigorously for 10–15 times.
3. Immediately transfer the suspension into a 1.5 mL RNase-free Eppendorf tube. Incubate on a nutator or rotating wheel for 5 min at 4 °C.
4. Pellet the chromatin by spinning at $1000 \times g$ for 3 min at 4 °C. Carefully remove the supernatant using pipette instead to vacuum to avoid losing the chromatin pellet (*see Note 4*).
5. Resuspend the chromatin pellet carefully with 1 mL of 1× NUN buffer. Incubate on a nutator or rotating wheel for 5 min at 4 °C. Remove the supernatant carefully with pipette. Repeat this wash twice more. For the last spinning, use $5000 \times g$ for 5 min at 4 °C (*see Note 5*).
6. Add 1 mL of TRIzol Reagent into chromatin pellet and vortex for 30 s. The chromatin pellet should be still visible.

7. Incubate the homogenized sample for 5 min at 50 °C to permit complete dissolving of chromatin pellet and dissociation of the nucleoprotein complex vortex for 30 s.
8. Process the rest steps of RNA extraction following standard protocol of RNA purification using TRIzol. Resuspend the RNA pellet in 85 μ L of RNase-free water.
9. Add 10 μ L of 10 \times DNase I Reaction Buffer and 5 μ L of DNase I (RNase-free) into RNA solution and mix well by pipette. Leave at room temperature for 20 min.
10. Add 300 μ L of RNase-free water into RNA solution.
11. Add 200 μ L of chloroform into RNA solution. Process the rest steps of RNA extraction following standard protocol of RNA purification using TRIzol. Resuspend the RNA pellet in 40 μ L of RNase-free water. Leave the RNA solution on ice.
12. Measure the concentration of purified RNA with NanoDrop. A260/A280 is \sim 2.0. Low A260/A280 values (<1.9) often suggest DNA contamination. Normally, 0.5–4 μ g RNA is obtained. The nascent RNA solution can be preceded into library preparation directly or stored at -80 °C until use.
13. (*Optional*) Verify the quality of nascent RNA by measuring the level of intron retention. Generate cDNA library with \sim 0.5 μ g nascent RNA to compare with cDNA made from total RNA. Use quantitative real-time PCR to measure the level of exon and intron for specific genes (Fig. 1).

3.4 Preparation of Nascent RNA Library

1. Check nascent RNA quality with either the Agilent RNA 6000 Nano Kit (for 5–500 ng/ μ L of starting material) or Pico Kit (for 50–5000 pg/ μ L of starting material) on the 2100 Agilent Bioanalyzer. Normally, the average size of nascent RNA is larger than total RNA due to the lack of co-transcriptional RNA splicing. And the amount of rRNA in nascent RNA is relatively less than that in total RNA (Fig. 2) (*see Note 6*).

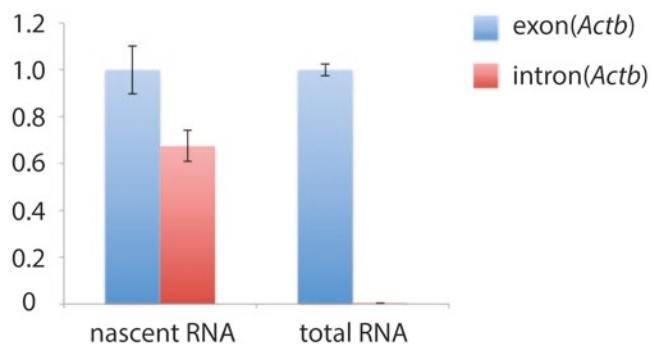


Fig. 1 The level of transcripts at exon and intron on *Actb* for nascent and total RNA in mouse embryonic stem cells. It is normalized on the level of exon

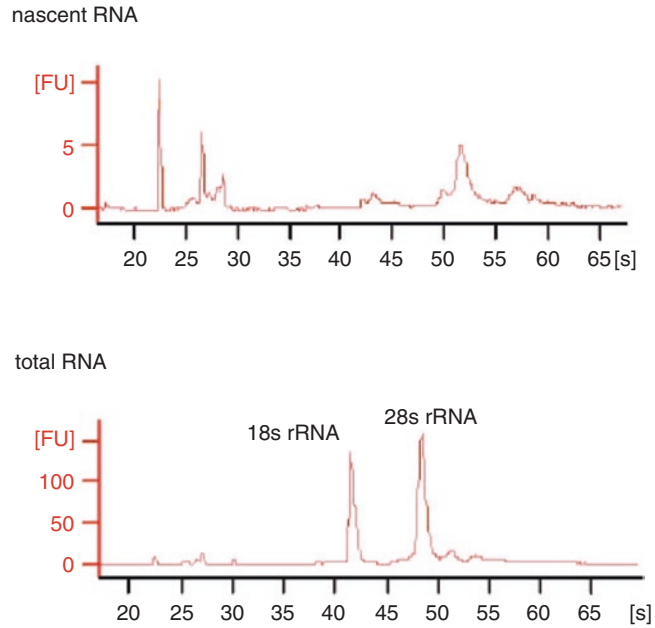


Fig. 2 Bioanalyzer electropherogram of nascent and total RNA

2. Dilute 1 μg of nascent RNA to a total volume of 10 μL using RNase-free water. If the concentration is lower than 0.1 $\mu\text{g}/\mu\text{L}$, use 10 μL without dilution. And 0.1 μg of nascent RNA is considered the minimal total amount (*see Note 7*).
3. Deplete rRNA, synthesize cDNA and construct and amplify libraries as instructed by the TruSeq Stranded Total RNA Sample Preparation Guide (#15031048) (Illumina, San Diego, CA, USA).
4. Quantify completed library using Qubit 2.0 and Qubit dsDNA HS Assay Kit. About ~20–75 $\text{ng}/\mu\text{L}$ is expected with 1 μg of starting material.
5. Validate library quality with either the Agilent DNA 1000 Kit or the Agilent High Sensitivity DNA Kit on the 2100 Agilent Bioanalyzer. We expect smooth profiles peaking near 250 bp with no measurable primer dimer contamination (*see Note 8*).

3.5 Next-Generation Sequencing

1. Pool uniquely indexed libraries evenly. This is achieved using the measured library concentration (Section 3.4, Step 4) and the average bp size (Section 3.4, Step 5) to calculate the concentration (nM) of each sample. The libraries are pooled so as to have equal amount (fmol) of each sample.
2. Denature and dilute libraries as instructed in Sequencing documentation.
3. Sequence on an Illumina sequencer (*see Note 9*).

4 Notes

1. Cell number used for each nascent-seq experiment can be increased to 1×10^8 or more in order to capture the lowly expressed nascent transcripts.
2. $2\times$ NUN Buffer should be prepared with RNase-free water to avoid RNA degradation. It is recommended to prepare Nuclei Wash Buffer with RNase-free water too.
3. Keep the samples on ice as much as possible when processing nuclei isolation to preserve the integrity of DNA–RNA–RNA polymerase ternary complex.
4. As the chromatin pellet does not stick to the bottom of the tube after spinning, use pipette rather than vacuum to carefully remove the supernatant.
5. The chromatin pellet is barely dissolved in TRIzol at room temperature. Incubate the chromatin pellet at 50°C for ~ 5 min for a complete dissolving of chromatin.
6. The amount of rRNA in nascent RNA compared with that in total RNA can be used to evaluate the purity of nascent RNA. For purified nascent RNA with total RNA contamination, the level of processed rRNA is relatively higher.
7. DNA digestion by DNase I is required as the DNA content is constantly high in the RNA solution purified from chromatin. If the A260/A280 value is lower than 1.9 after DNase I treatment, another round of DNA digestion is recommended.
8. cDNA library of high quality should have no visible primer dimers, which form clusters efficiently and take valuable space on the flow cell without generating any useful data.
9. Libraries must be sequenced on an Illumina sequencer since the Illumina TruSeq Stranded Total RNA Sample Preparation kit is used to prepare the libraries.

References

1. Holoch D, Moazed D (2015) RNA-mediated epigenetic regulation of gene expression. *Nat Rev Genet* 16:71–84. <https://doi.org/10.1038/nrg3863>
2. Orom UA, Shiekhattar R (2013) Long noncoding RNAs usher in a new era in the biology of enhancers. *Cell* 154:1190–1193. <https://doi.org/10.1016/j.cell.2013.08.028>
3. Smith E, Shilatifard A (2013) Transcriptional elongation checkpoint control in development and disease. *Genes Dev* 27:1079–1088. <https://doi.org/10.1101/gad.215137.113>
4. Liu X, Bushnell DA, Kornberg RD (2013) RNA polymerase II transcription: structure and mechanism. *Biochim Biophys Acta* 1829:2–8. <https://doi.org/10.1016/j.bbagr.2012.09.003>
5. Bentley DL (2014) Coupling mRNA processing with transcription in time and space. *Nat Rev Genet* 15:163–175. <https://doi.org/10.1038/nrg3662>
6. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*

- 322:1845–1848. <https://doi.org/10.1126/science.1162228>
7. Kwak H, Fuda NJ, Core LJ, Lis JT (2013) Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* 339:950–953. <https://doi.org/10.1126/science.1229386>
 8. Churchman LS, Weissman JS (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* 469:368–373. <https://doi.org/10.1038/nature09652>
 9. Mayer A, di Iulio J, Maleri S, Eser U, Vierstra J, Reynolds A, Sandstrom R, Stamatoyannopoulos JA, Churchman LS (2015) Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* 161:541–554. <https://doi.org/10.1016/j.cell.2015.03.010>
 10. Nojima T, Gomes T, Grosso AR, Kimura H, Dye MJ, Dhir S, Carmo-Fonseca M, Proudfoot NJ (2015) Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* 161:526–540. <https://doi.org/10.1016/j.cell.2015.03.027>
 11. Khodor YL, Rodriguez J, Abruzzi KC, Tang CH, Marr MT II, Rosbash M (2011) Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in *Drosophila*. *Genes Dev* 25:2502–2512. <https://doi.org/10.1101/gad.178962.111>
 12. Chen FX, Woodfin AR, Gardini A, Rickels RA, Marshall SA, Smith ER, Shiekhatter R, Shilatifard A (2015) PAF1, a molecular regulator of promoter-proximal pausing by RNA polymerase II. *Cell* 162:1003–1015. <https://doi.org/10.1016/j.cell.2015.07.042>
 13. Chen F, Gao X, Shilatifard A (2015) Stably paused genes revealed through inhibition of transcription initiation by the TFIIF inhibitor triptolide. *Genes Dev* 29:39–47. <https://doi.org/10.1101/gad.246173.114>
 14. Liang K, Woodfin AR, Slaughter BD, Unruh JR, Box AC, Rickels RA, Gao X, Haug JS, Jaspersen SL, Shilatifard A (2015) Mitotic transcriptional activation: clearance of actively engaged pol II via transcriptional elongation control in mitosis. *Mol Cell* 60:435–445. <https://doi.org/10.1016/j.molcel.2015.09.021>
 15. Cai H, Luse DS (1987) Transcription initiation by RNA polymerase II in vitro. Properties of preinitiation, initiation, and elongation complexes. *J Biol Chem* 262:298–304
 16. Wuarin J, Schibler U (1994) Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Mol Cell Biol* 14:7219–7225

Genome-Wide Copy Number Alteration Detection in Preimplantation Genetic Diagnosis

Lieselot Deleye, Dieter De Coninck, Dieter Deforce,
and Filip Van Nieuwerburgh

Abstract

Shallow whole genome sequencing has recently been introduced for genome-wide detection of chromosomal copy number alterations (CNAs) in preimplantation genetic diagnosis (PGD), using only 4–7 trophoctoderm cells biopsied from day-5 embryos. This chapter describes the complete method, starting from whole genome amplification (WGA) on isolated blastomere(s), up to data analysis for CNA detection. The process is described generically and can also be used to perform CNA analysis on a limited number of cells (down to a single cell) in other applications. This unique description also includes some tips and tricks to increase the chance of success.

Key words Massive parallel sequencing (MPS), Shallow whole genome sequencing, Preimplantation genetic diagnosis (PGD), Whole genome amplification (WGA), Copy number alterations (CNAs)

1 Introduction

Massively parallel sequencing (MPS)-based preimplantation genetic diagnosis (PGD) has been the subject of several studies in recent years [1–4]. Those studies show the advantages of MPS over current methods, such as array comparative genomic hybridization (arrayCGH), for detecting chromosomal aberrations in PGD [1–3]. Although arrayCGH proved its value, its rather limited resolution and the relative high cost are a disadvantage [3]. Shallow or low-pass whole genome sequencing can address these issues [1]. Deleye et al. (2015) concluded that shallow whole genome sequencing on trophoctoderm biopsies is a preferable alternative for the detection of chromosomal structural and numerical abnormalities in PGD embryos [1]. MPS-based PGD was able to detect chromosomal aberrations equal to or larger than 3 Mb in 47 blastocysts of 15 patients with a better resolution and signal/noise ratio compared to arrayCGH-based PGD [1]. Although no large clinical trials on the long-term clinical advantages of embryo

selection using MPS have been reported, a few cases of birth of a healthy baby are known [3, 5].

Embryo implantation fitness is determined using only 4–7 trophoblast cells biopsied from day-5 embryos [6–8]. Whole genome amplification (WGA) is needed to amplify the DNA from those cells before downstream analysis. Especially with such low amounts of input DNA, some WGA methods will lead to unbalanced amplification with over- and under-representation of genomic regions. Bias introduced during this amplification process may lead to misinterpretations of the genomic profile [9]. Choosing the correct method for amplification depends on the application [10]: PCR-based methods are better suited for chromosomal aberration detection compared to multiple displacement amplification (MDA) methods, because they give a more balanced genomic amplification. Recently, two state-of-the-art PCR-based WGA methods were compared to study their applicability for copy number alteration (CNA) detection using MPS [9]. In this study, Picoplex/SurePlex (Rubicon Genomics Inc., MI 48108, USA/BlueGnome Ltd., Mill Court, Great Shelford, Cambridge, UK) proved to be better suited for CNA analysis using MPS compared to Multiple Annealing and Looping Based Amplification Cycles (MALBAC) (Yikon genomics, Beijing, China): SurePlex WGA is more uniform across the genome, leading to less false positive and false negative CNAs. SurePlex WGA has been successfully applied in clinical MPS-based PGD with correct detection of CNAs with a resolution of 3 MB [1].

CNA analysis starting from a limited amount of DNA has several applications. Therefore, a detailed description of this method is useful. In this chapter, the method is described in detail as it would be executed for PGD. The general techniques described are from standard protocols, but important changes to these standard protocols were introduced to optimize the results. The note section, listing some tips and tricks, is important to increase the chance of success. The complete procedure, starting from a few cells and ending up with a CNV profile of the DNA, has never been described in such great detail before. The method, as described, has proven its success in detecting CNV up to 3 Mb starting from 4 to 6 blastocysts.

The protocol starts with WGA on a few cells. Based on previous results, the SurePlex WGA kit was the method of choice for WGA [9]. SurePlex might be replaced with another WGA method, but this might have a significant influence on the results. The genomic coverage, sequence error rate, yield, and representation bias might differ, possibly leading to less accurate CNA detection and/or a lower CNA detection resolution. Changing from SurePlex to PicoPlex should not influence the results, since both the kits basically have the same underlying method. The amplified DNA is fragmented to fragments of 200 bp using sonication. Subsequently, Illumina sequencing libraries are prepared from the fragmented DNA using NEBNext Ultra 2 library preparation kit.

This kit is suited for DNA input amounts as low as 500 pg. The needed library preparation method depends on the downstream sequencing technology. In this protocol, Illumina NextSeq500 sequencing is described. Ion Torrent sequencing will show similar results as demonstrated in a previous report [1].

2 Materials

2.1 Whole Genome Amplification

1. DNA Lo-bind 0.2 mL or 1.5 mL tubes.
2. Filter tips.
3. Positive control (genomic DNA at a known concentration): diluted to 30 pg/ μ L in molecular biology grade water (*see Note 1*).
4. Phosphate-Buffered Saline (PBS).
5. SurePlex/PicoPlex WGA kit for single-cell whole genome amplification, store at -20°C (Rubicon Genomics) (*see Note 2*).
6. Thermocycler with heated lid at 105°C .
7. Purification kit: Genomic Clean & Concentrator (Zymo Research, Irvine, CA, USA).
8. Benchtop microcentrifuge.
9. Thermomixer at 65°C .
10. Quality control using capillary gel electrophoresis: High-Sensitivity DNA kit (lab-chip), 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).
11. Concentration measurement: Qubit[®] dsDNA High Sensitivity Assay kit (ThermoFisher Scientific, Waltham, MA, USA).

2.2 Fragmentation of Purified WGA Product

1. DNA Lo-bind 0.2 mL tubes.
2. Filter tips.
3. Fragmentation of DNA by sonication: S2 Focused Ultrasonicator with Adaptive Focused Acoustics technology and MicroTUBES (Covaris, Woburn, MA, USA).
4. $1/5\times$ Tris-ethylenediaminetetraacetic acid (EDTA) buffer: 0.5 mL $20\times$ TE buffer, 49.5 mL molecular biology grade water.

2.3 Library Preparation

1. DNA Lo-bind 0.2 mL or 1.5 mL tubes.
2. Filter tips.
3. All buffers mentioned during the downstream protocol are derived from the NEBNext Ultra II kit. Store at -20°C . (New England Biolabs, Ipswich, MA, USA) (*see Note 3*).
4. Thermocycler with adjustable heated lid.
5. Thermomixer for 0.2 and 1.5 mL tubes.

6. Purification kit: Genomic Clean & Concentrator (Zymo Research).
7. Benchtop microcentrifuge.
8. Size selection: E-Gel 2% EX agarose gel, 1 Kb plus DNA ladder and E-gel ibase power system (ThermoFisher Scientific).
9. Purification from gel: Zymoclean gel DNA recovery kit (Zymo Research).
10. Dark Reader blue light Transilluminator.
11. 10 ng/ μ L yeast tRNA.
12. Purification after enrichment PCR: magnetic beads, AMPure XP beads (Beckman Coulter, Carlsbad, CA, USA) and magnetic particle concentrator (MPC).
13. 80% ethanol: 40 mL 100% ethanol with 10 mL molecular bio-grade water. This can be kept at 4 °C for 2 weeks.

2.4 Library Quality Control and Quantification

1. Quality control using capillary gel electrophoresis: High-Sensitivity DNA kit (lab-chip) and 2100 Bioanalyzer (Agilent Technologies).
2. Quantification of adapter-ligated fragments: Sequencing library qPCR quantification guide (Illumina, San Diego, CA, USA).
3. Sequencing on an Illumina instrument (*see Note 4*).

3 Methods

Carry out all the procedures at room temperature unless otherwise specified.

3.1 Whole Genome Amplification

1. Isolate the necessary amount of cells in less than 2.5 μ L of cell medium or PBS in a 0.2 mL tube (*see Note 5*).
2. Dilute with the appropriate volume of Cell Extraction Buffer to achieve a total sample volume of 5 μ L (*see Note 6*).
3. Prepare a positive and negative control. Take 2.5 μ L of PBS as negative control and take 2.5 μ L of the diluted positive control.
4. Prepare an extraction cocktail for at least five samples (positive and negative control included) (*see Note 7*). Combine 24 μ L of Extraction Enzyme Dilution Buffer and 1 μ L of Cell Extraction Enzyme in a 0.2 mL tube and mix by flicking the tube. Add 5 μ L of this cocktail to each sample (*see Note 8*).

Incubate the samples immediately in a pre-programmed thermocycler:

10 min	75 °C
4 min	95 °C
Hold	10 °C

Proceed immediately.

5. Prepare a pre-amp cocktail for at least five samples. Combine 24 μL of Pre-Amp buffer and 1 μL of Pre-Amp Enzyme in a 0.2 mL tube and mix by flicking the tube. Add 5 μL of the cocktail to each sample (*see Note 8*). Incubate the samples in a pre-programmed thermocycler:

2 min	95 °C	
15 s	95 °C	} 12 cycles
50 s	15 °C	
40 s	25 °C	
30 s	35 °C	
40 s	65 °C	
40 s	75 °C	
Hold	4 °C	

6. Place the samples on ice before use.
7. Prepare the Amplification Cocktail as instructed. Combine 34.2 μL nuclease-free water/sample, 25 μL Amplification buffer/sample and 0.8 μL Amplification enzyme/sample in a 0.2 mL tube. Mix by flicking the tube and add 60 μL to each sample tube. Incubate the samples in a pre-programmed thermocycler:

2 min	95 °C	
15 s	95 °C	} 14 cycles
1 min	65 °C	
1 min	75 °C	
Hold	4 °C	

8. Vortex the samples and spin down.
9. Purify the samples on spin columns. As an efficient example, Zymo Genomic Clean and Concentrator, is described below.
10. Place nuclease-free water at 65 °C in a thermomixer.

11. Add 375 μL Binding buffer to each sample and transfer this mix to the spin column (*see Note 9*).
12. Centrifuge in a microfuge at 11,000 rpm ($11,092 \times g$) for 25 s.
13. Discard the flow-through and pat the collector tube dry. Re-use the same collector tube.
14. Wash the spin columns with 200 μL Wash buffer.
15. Centrifuge in a microfuge at 11,000 rpm ($11,092 \times g$) for 25 s.
16. Repeat **step 14**.
17. Centrifuge in a microfuge at 11,000 rpm ($11,092 \times g$) for 25 s (*see Note 10*).
18. Transfer the spin columns to a new 1.5 mL tube.
19. Elute the DNA in 32 μL of the pre-warmed molecular biology grade water (*see Note 11*) and incubate for 1 min at room temperature.
20. Centrifuge in a microfuge at 11,000 rpm ($11,092 \times g$) for 35 s.
21. Remove the spin column and store the 1.5 mL tubes at $-20\text{ }^{\circ}\text{C}$ or perform a quality check before storing. Measure the concentration with Qubit (*see Note 12*). Analyze the sample on a high sense Agilent lab-chip (*see Note 13*).

3.2 Fragmentation of Purified WGA Product

1. Dilute 100 ng of the WGA product with $1/5 \times \text{TE}$ buffer in 130 μL in microTUBES.
2. Adjust the settings of the Covaris S2 for a 200 bp fragmentation: duty cycle of 10%, intensity of 5, 200 cycles/burst, and a treatment time of 190 s (*see Note 14*).

3.3 Library Preparation

1. Make sure the heated lid of the thermocycler is set on $75\text{ }^{\circ}\text{C}$.
2. Add 7 μL of End repair reaction buffer to empty 0.2 mL tubes. Transfer 50 μL of each fragmented sample into a 0.2 mL tube with buffer. Add 3 μL of End prep enzyme mix to each sample and mix by flicking the tube. Incubate the samples immediately in the pre-programmed thermocycler with heated lid on $75\text{ }^{\circ}\text{C}$:

30 min	$20\text{ }^{\circ}\text{C}$
30 min	$65\text{ }^{\circ}\text{C}$
Hold	$4\text{ }^{\circ}\text{C}$

Proceed immediately.

During incubation, put a thermomixer at $20\text{ }^{\circ}\text{C}$ and pre-program another thermocycler at $37\text{ }^{\circ}\text{C}$ with a heated lid at $47\text{ }^{\circ}\text{C}$. Make sure the temperature of the thermocycler is already at $37\text{ }^{\circ}\text{C}$ when the samples are introduced. If DNA input was less than 100 ng, dilute the adapter to a final concentration of $1.5\text{ }\mu\text{M}$ (*see Note 15*).

3. Add 30 μL ligation master mix, 1 μL ligation enhancer and 2.5 μL diluted adapter to each sample in this respective order (*see Note 16*). Mix by flicking the tube and incubate for 15 min at 20 $^{\circ}\text{C}$ in a thermomixer.
4. Add 3 μL USER enzyme to each sample and incubate 15 min at 37 $^{\circ}\text{C}$ in a thermocycler with heated lid at 47 $^{\circ}\text{C}$.
5. Purify the samples on spin columns, such as described above. Place nuclease-free water at 65 $^{\circ}\text{C}$ in a thermomixer.
6. Add 482.5 μL Binding buffer to each sample and transfer this mix to a spin column (*see Note 9*).
7. Follow **steps 12–18** as in Subheading **3.1**.
8. Elute the DNA in 22 μL of the pre-warmed nuclease-free water (*see Note 17*) and incubate for 1 min at room temperature.
9. Centrifuge in a microfuge at 11,000 rpm (11,092 $\times g$) for 35 s and remove the spin column (*see Note 18*).
10. Perform size selection using E-gel EX 2% agarose gels. Dilute the DNA ladder $\frac{1}{4}$ in 20 μL water. Add all samples individually to a gel-slot, alternated with a ladder every three or four samples. Choose the 1–2% gel program on the iBase (10 min) and run the gel.
11. Remove the gel from its case and visualize using a Dark Reader blue light transilluminator. Cut the gel at the desired height to retrieve the DNA (*see Note 19*).
12. Dissolve the gel in 300 μL ADB buffer (Zymo gel purification) at 55 $^{\circ}\text{C}$ in a thermomixer for at least 10 min. When the gel is completely dissolved, purify on spin columns as described above in Subheading **3.1**. However, change the elution volume to 17 μL .
13. Next, enrich the samples carrying adapters. Each sample is assigned an index (*see Note 20*). Transfer the samples to a 0.2 mL tube and add 1 μL of tRNA (*see Note 21*). Add 3 μL of the assigned index primer and 3 μL of universal primer to each sample. Finally, transfer 25 μL of HF 2 \times PCR master mix to each sample. Mix by flicking the tube and immediately incubate in a pre-programmed thermocycler:

30 s	98 $^{\circ}\text{C}$		
10 s	98 $^{\circ}\text{C}$	}	9 cycles (<i>see Note 22</i>)
75 s	65 $^{\circ}\text{C}$		
5 min	65 $^{\circ}\text{C}$		
Hold	4 $^{\circ}\text{C}$		

While waiting: put the magnetic beads at room temperature.

14. Purify the sample using magnetic beads (*see Note 23*). Mix the samples with 45 μ L beads and leave at room temperature for 6 min.
15. Put on a MPC until the liquid is clear and remove the supernatant.
16. Wash the beads with 200 μ L 80% ethanol while on MPC and wait 30 s before removing the supernatant.
17. Repeat **step 16** above.
18. Make sure all supernatant has been removed, in order to facilitate the air-drying process. Air-dry the samples until cracks are visible between the beads (*see Note 24*).
19. Add 22 μ L nuclease-free water to each sample and remove from MPC. Mix well until sample and beads are a homogeneous mixture. Leave at room temperature for 3 min.
20. Put on MPC until clear and transfer the supernatants to new 0.2 mL tubes. Make sure no beads are transferred.

3.4 Library Quality Control and Quantification

Library quality control is performed by capillary gel electrophoresis on a Bioanalyzer High sensitivity lab-chip. The result is displayed as an electropherogram (EPG) showing intensity in function of fragment-size distribution. The library should contain DNA fragments of around 300 bp. Figure 1a shows an EPG of a good quality library. The other two smaller peaks represent the internal standards (upper and lower marker). The intensity of the library informs about the concentration of the library (*see Note 25*). A peak visible around 85 bp, as the one shown in Fig. 1b, indicates the presence of primer-dimers (*see Note 26*).

The quantification of adapter-ligated fragments in the library is performed by qPCR. Only fragments carrying an adapter will bind to the flowcell, and therefore will be sequenced. qPCR is performed as recommended by Illumina. Make sure the libraries are diluted to fit the standard curve used for qPCR. The concentration measured after qPCR reflects the amount of DNA that can be sequenced (*see Note 27*).

The libraries are further prepared for sequencing using the standard Illumina protocols. Sequencing is performed on a NextSeq500 using a high output flowcell for single read and 75 cycles. The sequencing run will last for about 11 h.

3.5 Data Analysis

Several tools and programs are available to detect CNAs using shallow, genome-wide massively parallel sequencing data [11, 12]. However, most of these tools only handle a certain part of the analysis. It requires bioinformatics knowledge to handle and combine these different tools to perform CNA detection starting from raw

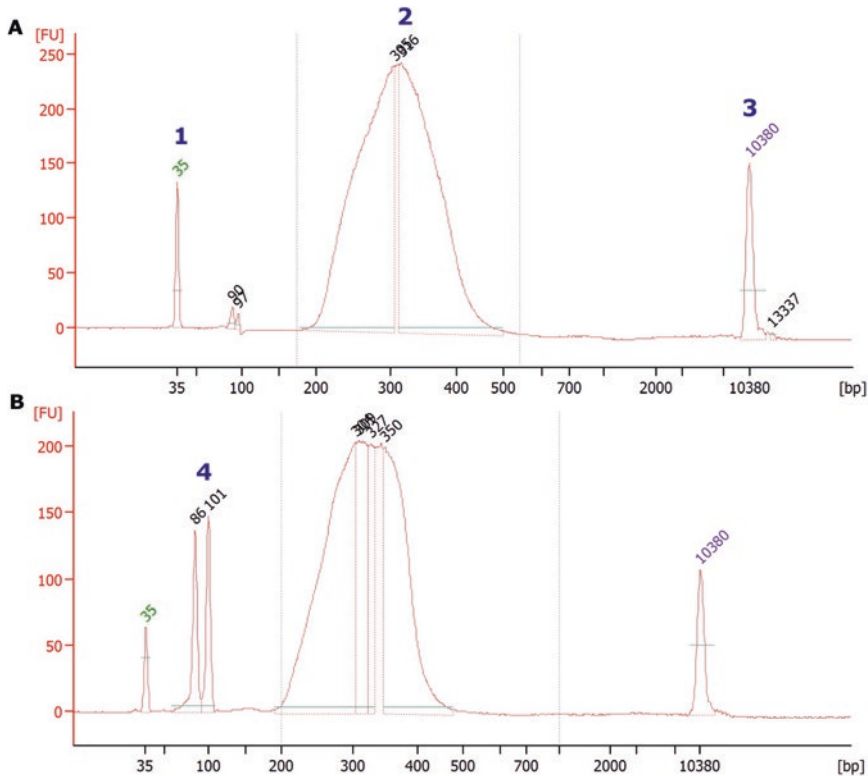


Fig. 1 Library quality control. (a) An electropherogram of a good quality library with a fragment length of ± 300 bp. (1) Lower marker, (2) Library, (3) Upper marker. (b) An electropherogram of a library with primer-dimers (4)

sequencing data. Recently, ViVar [13] was developed to provide a user-friendly web-based analysis platform that handles all necessary steps in a comprehensive way. This platform is freely available at <https://www.cmgg.be/vivar/>. Download and installation instructions are also provided at this web URL.

ViVar offers an easy-to-use and straightforward interface which enables the analysis from raw sequencing data, obtained from the sequencer as .fastq files, for the detection of CNAs:

1. Select “Projects” from the top navigation bar (Fig. 2).
2. Click the “New project” button to create a new project (Fig. 2).
3. Click the “Save” button after filling out the project’s information to save the new project.
4. One can now find the new project in the list, which is accessed by selecting “Projects” from the top navigation bar (Fig. 2).
5. By clicking on the new project’s name, a page containing the experiments for that project will be loaded. In ViVar, each experiment consists of a single sample.

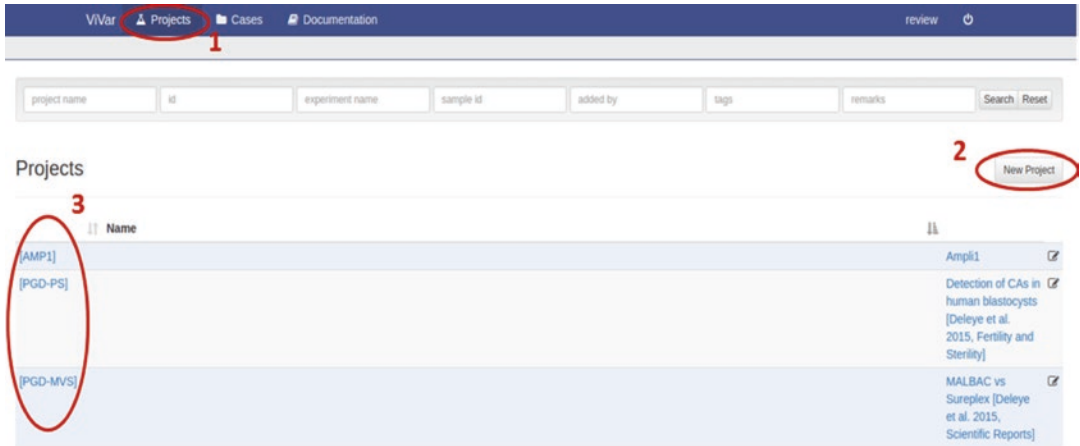


Fig. 2 Creating a new project in ViVar. First, select 'Projects' from the top navigation bar (1), then click the 'New project' button to create a new project (2). After filling out the project's information, the newly created project will appear in the projects list (3)

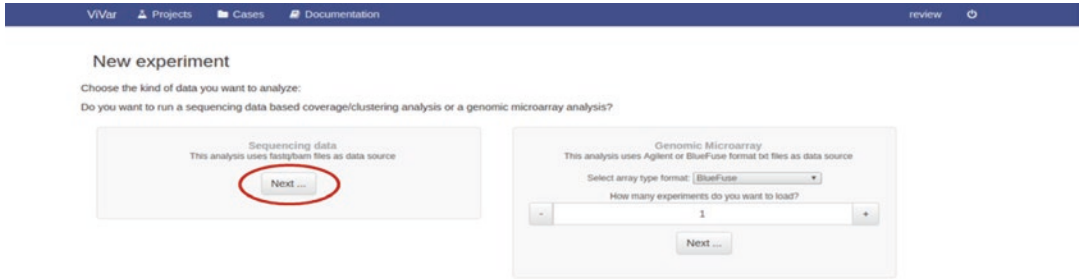


Fig. 3 Starting a new sequencing experiment in ViVar. A new experiment/analysis based on massively parallel sequencing data can be started by clicking the 'Next' button in the 'Sequencing data' field (red circle)

6. To create a new experiment, click the “New experiment” button.
7. Choose “Sequencing data” by clicking the “Next” button in the “Sequencing data” field (Fig. 3).
8. From the drop-down list select the project you want to add a new experiment to (Fig. 4).
9. Add data to an experiment by selecting the samples you want to analyze. You can select multiple samples at once (Fig. 4).
10. Next, choose the organism from which the samples originate (Fig. 4).
11. Then, also select the version of the reference genome you want to use for the analysis from the drop-down list (Fig. 4).
12. Finally, in the “Depth of Coverage” field, select the bin size you want to use for the analysis (see description of this parameter

New sequencing experiment

1. Project to add the experiments to:

Select a project

2. Add data

Select a file

Paste a path

Search:

Name	Library Type	Files
47LS-C3	Library type: Single reads	47LS-C3.fastq.gz

3. Choose the organism

Homo sapiens

Choose the desired genome build

GRCh37

4. Choose from following analyses

Options available: readDepth, bowtie, coverage

readDepth analysis of coverage
Load coverage windows of readDepth analysis

OFF

Bowtie2 read alignment
Read sequencing reads using Bowtie v2 read alignment, do a bam file sort and index.

OFF

Depth of coverage
Depth of coverage/ QDNaseq analysis

1 Select binsize: QDNaseq GRCh37.1000kbp.SR50

2 OFF

3

Fig. 4 Setting the parameters for a new analysis in ViVar. Once a new analysis based on massively parallel sequencing data has been created, parameters for the analysis can be specified: the project to which the experiment should be added, the sample(s) which should be analyzed, the reference genome which should be used in the analysis, and the bin size (1). Analysis can be started by switching the tumble button from 'OFF' to 'ON' (2) and clicking the 'Next...' button (3)

below), switch the tumble button from "OFF" to "ON" and click the "Next..." button to start the analysis (Fig. 4).

ViVar will now analyze the data in a fully automated manner. Briefly, it first uses Bowtie [14] to place the sequencing reads onto the reference genome. Then, using these mapped reads, it performs CNA detection using the QDNaseq algorithm [15]. To this end, the genome is divided into nonoverlapping fixed size parts, so called bins. The size of these bins will determine the minimum size of the CNAs which can be detected. Generally, CNAs four to five times the size of the bins can be detected. The number of reads mapped to each bin will be determined. This number is influenced by certain factors, such as GC-content and mappability, for which the number of reads mapped per bin is normalized. After GC-content and mappability

normalization, read counts are median-normalized by dividing the number of reads in each bin by the median number of reads across all the bins. As CNAs are assumed to be rare, the median number of reads across all the bins is a fair estimate of the expected number of reads per window for a perfectly diploid genome. As such, the median-normalized read counts represent a measure for the deviation from diploidy for each window and a copy number (CN) estimate is calculated using the following formula: $CN = 2 \cdot (\text{read count} / \text{median read count})$. Then, a circular binary segmentation (CBS) algorithm [16] is applied which groups bins into larger contiguous regions with an equal CN. The mean of read counts of the windows contained in the segments is used as an estimator of the copy number of the whole segment. After this segmentation, CNAs are called when the segment's $\log_2(CN/2)$ surpasses a certain threshold. This threshold can be specified by the user in ViVar (see further). Based on literature review and own experience a threshold of approx. 0.35 performs well.

After the analysis has been finished, ViVar offers some powerful visualizations of the data: line profile plots, karyo plots, genome heatmaps, etc. (Fig. 5). The procedure to obtain these visualizations is always similar:

1. Select “Projects” from the top navigation bar.
2. Click on the project of interest. A list of experiments within the project is now shown.
3. Select the experiment you want to visualize by ticking the checkbox in front of the experiments name (Fig. 6).
4. Select, from the dropdown list at the top of the screen, the visualization of interest.

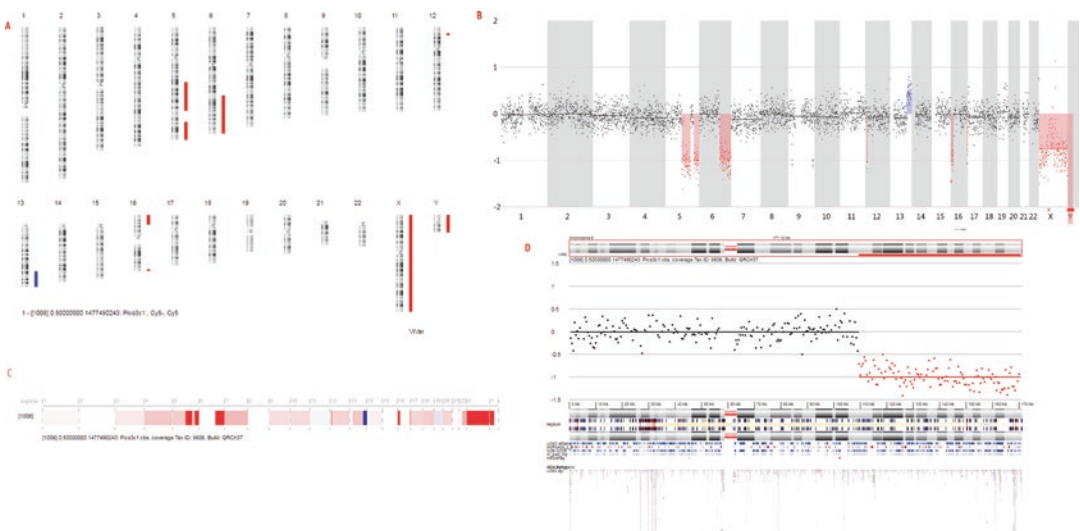


Fig. 5 Some of the visualizations possible with ViVar. Karyo plot (a); line view plot (b); genome heatmap (c); and chromosome view (d)

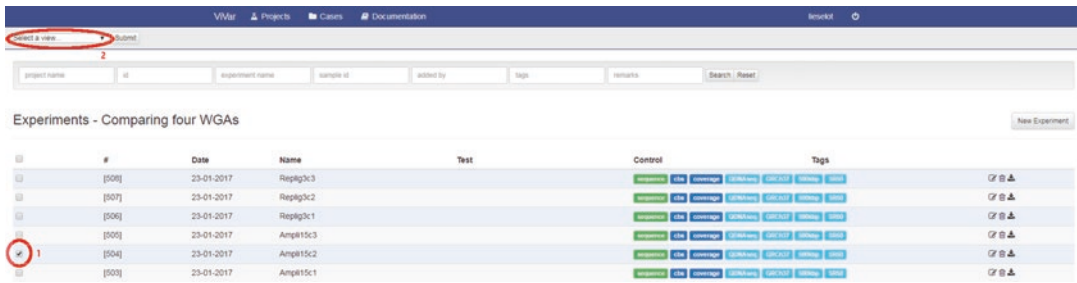


Fig. 6 Visualizing results in ViVar. Results can be visualized in ViVar by ticking the checkbox in front of the name of the experiment you want to visualize (1) and choosing the visualization of interest from the dropdown list (2)

5. Some visualizations, such as the karyo plots, can handle multiple simultaneous experiments.
6. Finally, certain visualization and analysis settings, such as color schemes and thresholds for calling CNAs can be adjusted by clicking on “criteria” in the upper right corner of the visualization screen.

4 Notes

1. The amount of DNA in the positive control is equivalent to five cells.
2. All buffers used during the downstream protocol are derived from this kit. Buffers must be vortexed before use, but never vortex enzymes. Always keep enzymes on ice.
3. Another library preparation method for Illumina, without the enrichment PCR step, is a valid alternative for the described method. Omitting the PCR-step will lead to a more uniform coverage of the genome [9]. However, this TruSeq PCR-free library prep kit requires more input material, which might pose a limitation for some applications. Replacing the NEBNext Ultra II kit with the NEBNext Ultra I version will not change the outcome.
4. Illumina sequencing kit: depends on the number of samples and the coverage aimed. Here, Illumina NextSeq500 high output kit v2 (75 cycles).
5. (a) Blastocysts are kept at -20°C after biopsy until genetic analyses. Cells isolated from a cell culture are snap frozen in N_2 immediately after collection and stored at -80°C until further use. (b) Other start material, such as fixed or microdissected cells, might lead to different results and might need some optimization.
6. This mix can be stored at -80°C . However, it is recommended to proceed immediately.

7. To avoid pipetting errors when pipetting very small volumes, prepare the mix for at least five samples.
8. To avoid the loss of cell material, do not touch the liquid already present in the tube while adding the cocktail.
9. To increase binding of the DNA, add 5× Binding buffer (ratio 1:5).
10. Make sure no residual wash buffer/ethanol is left, because this might lead to a suboptimal elution of the DNA. Centrifuge again if necessary.
11. Elution with *pre-warmed* water will lead to a better elution of the DNA. The elution volume depends on the downstream application. Make sure to get rid of the air-bubbles in your tip that arise because of the temperature difference. Pipet the water straight on the top of the column.
12. Other assays based on fluorescent DNA stains can also be used to measure the concentration. Spectrophotometric techniques such as Nanodrop are less accurate because the UV signal is not specific to DNA.
13. If the negative control shows a similar result on Qubit and Agilent as the samples, some contamination might have occurred during the WGA procedure. In this case, the samples might contain more amplified contaminated material than the amplified template. If both the samples and positive control show negative results, the PCR reaction during WGA might have failed. If the positive control looks fine, but one (or more) of the samples is negative, that specific sample probably lacked template at the start.
14. The treatment time might need to be slightly adjusted depending on the specific instrument and water bath temperature.
15. The dilution needs to be made fresh.
16. Do not mix the three components in advance, since this might create adapter-dimers. The adapter ligation could be suboptimal, since a large part of the adapters are already ligated to each other. Add the three components sequentially to one sample, then add the three components sequentially to the next sample, etc. Avoid a large time difference between the first and last samples, since the reaction already starts at room temperature.
17. The elution volume is important for the next step.
18. At this point, the sample could be stored at $-20\text{ }^{\circ}\text{C}$. However, long time storage might negatively affect the concentration of library with intact sequencing adapters.
19. Make sure to change scalpel between each sample. If DNA concentrations are low, the samples might be nearly invisible under the dark reader. Retrieve the samples between 200 and 400 bp.

20. Make sure each sample is assigned a unique index. The specific combination of indexes is only important if less than seven samples are pooled (see kit documentation). Dual-indexing is used when more than 24 samples are pooled.
21. In samples where template DNA concentrations are rather low, tRNA is added as a carrier that will adsorb to most of the tube wall, resulting in a more concentrated template DNA in the rest of the tube. Higher concentration increases the efficiency of primer annealing to the template. More template will be efficiently amplified and the amount of primer-dimers will decrease.
22. The number of cycles depends on the input amount of DNA. Decrease the number of cycles if possible, to decrease amplification bias.
23. Make sure the beads are well mixed before use. When removing the supernatant, check your tip for beads. If beads are visible, transfer the liquid back into the right tube and wait again until the liquid is clear.
24. The protocol states not to dry the beads until they are cracked. Nevertheless, based on our experience, the elution efficiency increases when the ethanol is completely evaporated.
25. The intensity of both internal standards should be similar. If this is not the case, the concentration measurements are not reliable.
26. Intense primer-dimer peaks should be removed from the library. They will skew qPCR results, because they also contain the binding place for the qPCR primers and thus yield an unreliable quantification of adapter-ligated fragments. The dimers are removed by size selection on gel, as described before. The bright band of 200–400 bp is cut from the gel and the dimers are left behind nearly at the end of the gel. Primer-dimers might also be avoided by decreasing the primer input during enrichment PCR or decreasing the amount of samples prepared simultaneously. A peak at approx. 125 bp on the EPG indicates the presence of adapter-dimers. They will compete with the library-fragments for binding places on the flowcell. Only fragments carrying the correct adapters on both ends will bind to the flowcell. Measuring the amount of DNA from the library that will actually bind to the flowcell is essential to overcome over- or underclustering during the sequencing run and can only be achieved using qPCR.
27. When the fragment length between the samples and PhiX is different, a size correction is performed on the measured concentration after qPCR. PhiX has a fragment length of 500 bp, while the samples have fragment lengths of only 300 bp. For size correction, the measured concentration is multiplied by the ratio of PhiX length over library length.

References

- Deleye L, Dheedene A, De CD, Sante T, Christodoulou C, Heindryckx B et al (2015) Shallow whole genome sequencing is well suited for the detection of chromosomal aberrations in human blastocysts. *Fertil Steril* 104:1276–1285
- Fiorentino F, Biricik A, Bono S, Spizzichino L, Cotroneo E, Cottone G et al (2014) Development and validation of a next-generation sequencing-based protocol for 24-chromosome aneuploidy screening of embryos. *Fertil Steril* 101:1375–1382
- Wells D, Kaur K, Grifo J, Glassner M, Taylor JC, Fragouli E et al (2014) Clinical utilisation of a rapid low-pass whole genome sequencing technique for the diagnosis of aneuploidy in human embryos prior to implantation. *J Med Genet* 51:553–562
- Yin X, Tan K, Vajta G, Jiang H, Tan Y, Zhang C et al (2013) Massively parallel sequencing for chromosomal abnormality testing in trophectoderm cells of human blastocysts. *Biol Reprod* 88:69
- Tan Y, Yin X, Zhang S, Jiang H, Tan K, Li J et al (2014) Clinical outcome of preimplantation genetic diagnosis and screening using next generation sequencing. *Gigascience* 3:30
- De Vos A, Staessen C, De Rycke M, Verpoest W, Haentjens P, Devroey P et al (2009) Impact of cleavage-stage embryo biopsy in view of PGD on human blastocyst implantation: a prospective cohort of single embryo transfers. *Hum Reprod* 24:2988–2996
- Los FJ, Van Opstal D, van den Berg C (2004) The development of cytogenetically normal, abnormal and mosaic embryos: a theoretical model. *Hum Reprod Update* 10:79–94
- Vanneste E, Voet T, Le CC, Ampe M, Konings P, Melotte C et al (2009) Chromosome instability is common in human cleavage-stage embryos. *Nat Med* 15:577–583
- Deleye L, De Coninck D, Christodoulou C, Sante T, Dheedene A, Heindryckx B et al (2015) Whole genome amplification with SurePlex results in better copy number alteration detection using sequencing data compared to the MALBAC method. *Sci Rep* 5:11711
- de Bourcy CF, De Vlaminck I, Kanbar JN, Wang J, Gawad C, Quake SR (2014) A quantitative comparison of single-cell whole genome amplification methods. *PLoS One* 9:e105585
- Duan J, Zhang JG, Deng HW, Wang YP (2013) Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PLoS One* 8(3):e59128
- Zhao M, Wang Q, Wang Q, Jia P, Zhao Z (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* 14(Suppl 11):S1
- Sante T, Vergult S, Volders PJ, Kloosterman WP, Trooskens G, De Preter K et al (2014) ViVar: a comprehensive platform for the analysis and visualization of structural genomic variation. *PLoS One* 9:e113800
- Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25
- Scheinin I, Sic D, Bengtsson H, van de Wiel MA, Olshen AB, van Thuijl HF et al (2014) DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res* 24:2022–2032
- Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5:557–572

Multiplexed Targeted Sequencing for Oxford Nanopore MinION: A Detailed Library Preparation Procedure

Timokratis Karamitros and Gkikas Magiorkinis

Abstract

MinION is a small form factor sequencer recently retailed by Oxford Nanopore technologies. This lighter-sized USB3.0-interfaced device uses innovative nanotechnology to generate extra-long reads from libraries prepared using only standard molecular biology lab equipment. The flexibility and the portability of the platform makes it ideal for point-of-interest and real-time surveillance applications. However, MinION's limited capacity is not enough for the study of specific targets within larger genomes. Apart from just PCR-amplifying regions of interest, the capture of long reads spanning the edges of known-unknown genomic regions is of great importance for structural studies, such as the identification of mobile elements' integrations sites, bridging over low complexity repetitive regions etc.

In this study, using MinION-kit-included and commercially available reagents, we have developed an easy and versatile wet-lab procedure for the targeted enrichment of MinION libraries, capturing DNA fragments of interest before the ligation of the sensitive MinION sequencing-adapters. This method allows for simultaneous target-enrichment and barcode-multiplexing of up to 12 libraries, which can be loaded in the same sequencing run.

Key words MinION, Target-enrichment, Library preparation, Baits, Hybridization

1 Introduction

Membrane protein nanopores were first proposed as biosensors in the 90s' [1]. The idea was simple: as molecules pass through these microscopic holes, they disrupt an electric field applied to both the sides of an electrically resistant surface. This disruption could be characteristic of the identity of the molecule if there was an unambiguous way to pair them. It took almost 20 years of research and development and collaboration of multidisciplinary groups before numerous technical difficulties were resolved [2–4] and MinION, a USB3.0-interfaced, portable device, became one of Oxford Nanopore Technologies retail products.

MinION outputs electric current signatures, which are unique for each group of nucleotides that passes through each pore. These current signatures are uploaded to Oxford Nanopore cloud by a desktop

application, the EPI2ME operating system (provided by Metrichor company), which also returns the final reads after the online basecalling is completed. The HD5 formatted reads contain multiple types of information, including the usual nucleotide sequences and quality strings that can be easily transformed to fastq files with available tools [5]. The error rate of the raw MinION reads is high, with the identity of the raw 2D reads—double strand analyzed—ranging at approximately 92% compared to the reference. Several tools though have been recently described, which are capable of increasing the accuracy of MinION reads (*see Note 3*). Interestingly, variation calling with up to 99% accuracy is now feasible [6].

Given the limited total data output, the application of this promising technology on large genomes (e.g., human Whole Genome Sequencing) is impossible. A target enrichment method, which narrows the focus of MinION on specific genomic regions but also captures extra-long flanking sequences, is described in this chapter.

Single stranded oligos conjugated with biotin—known as “baits” (usually 60–120 bp long) to “fish” their compatible target-sequence of interest—are used in current target enrichment hybridization protocols [7, 8]. High read coverage over the targeted region is achieved after the removal of nonspecific, irrelevant DNA fragments [9, 10]. The fraction of the captured DNA is small and demands the post-capture amplification of the library, using primers against the pre-capture ligated sequencing adapters.

MinION sequencing adapters are not similar to those used in other platforms, which are just oligos. MinION adapters are special, both shape-wise and chemistry-wise. One the Y-shaped adapter is being ligated to the one end of each DNA fragment allowing the attachment of the first strand to the nanopore and a hairpin—U-shaped—adapter is being ligated to the other end of each DNA fragment, driving the second strand of the fragment into the nanopore, after the passage of the first has been completed. Both of them are platform-specific and are conjugated with special proteins, which saturate the DNA fragments on the sequencing chip and control their speed during their translocation through the nanopores (Fig. 1). Thus, the protein-nature of MinION sequencing adapters restricts their use with in-house developed protocols where the breaking of the biotin-streptavidin bond is needed [11].

Using the PCR-adapters included in the “Genomic DNA Sequencing kit –MAP 003” and the compatible barcoded primers we were able to capture and release the desired targets before the ligation of the sequencing protein-adapters (Fig. 2). We tested this method in two pilot experiments, targeting either a unique target within Lambda Phage genome alone, or all rRNA operon sequences in *E.coli* O157:H7 (str. Sakai) and *E.coli* W3110 (str. K12) genomes [12]. The bacterial libraries were pooled. A dedicated kit for multiplexing MinION libraries, the “PCR Barcoding Kit,” is

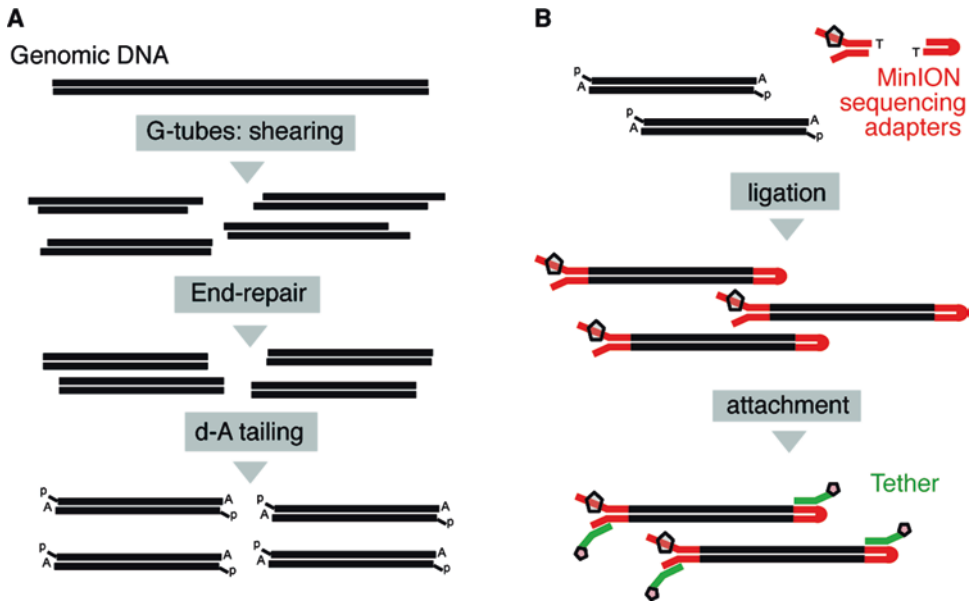


Fig. 1 Standard library-preparation for MinION genomic DNA sequencing. (a) Preprocessing of genomic DNA: Shearing is performed with g-tubes usually aiming in >8 Kbp fragments generation. The fragments are end-repaired and d-A tailed using standard library preparation modules. (b) Ligation of protein-conjugated (*rhombus*) MinION adapters (*red*) is followed by the attachment of the Tether (*green*), which anchors-saturate the fragments onto the nanopores flowcell. The library is conditioned and loaded to the MinION sequencer after this step

now available and the included PCR barcode adapters can be used instead of the standard ones—those for low-input DNA—for multiplexing the libraries in this protocol.

It is important to note that this method defers from just amplifying the targets using barcoded primers and following the library preparation for amplicon-sequencing, which would only amplify the intended sequences. Target enrichment allows the capture of sequences flanking the target itself, which is extremely useful for phasing and structural studies, especially when the flanks of the target are unknown: identification of viral and other mobile elements' integration sites, genomic rearrangements, large-scale insertions/deletions and bridging over repetitive regions are a few examples. This method, in conjunction with MinION's extra long reads, opens immense potential in this particular era, as we have shown that the capture of flanks more than 2000 bp long is feasible [12].

2 Materials

Materials needed but not included in Oxford Nanopore Technologies' kits:

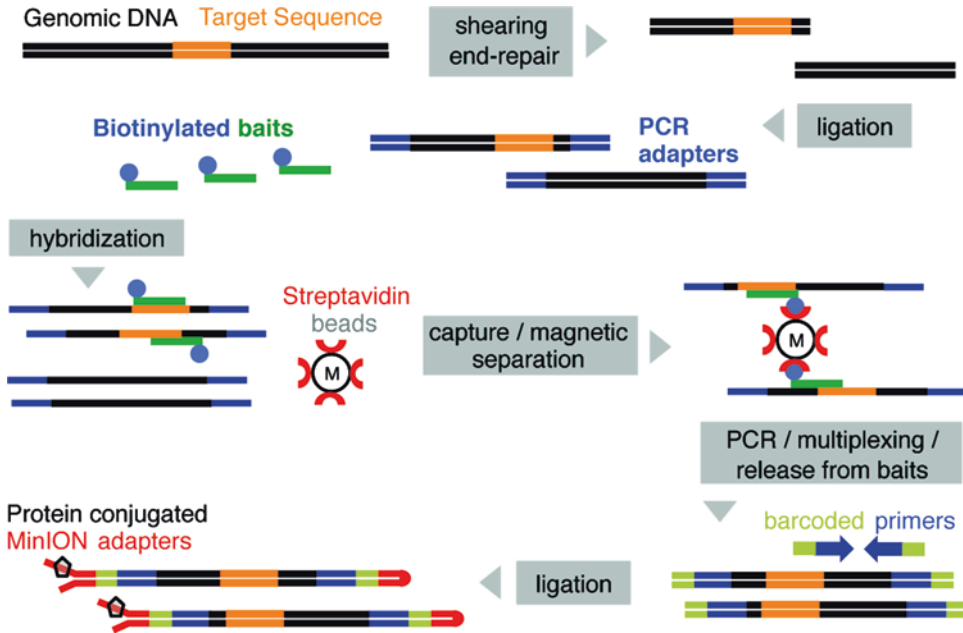


Fig. 2 The MinION target-enrichment library-preparation process. Genomic DNA (*black*) containing target sequences (*orange*) is sheared with G-tubes aiming in 5–6 Kbp fragments generation. PCR adapters (*blue*) are ligated to the end-repaired fragments. Biotinylated baits (*blue—green*) (optionally PCR-generated) are hybridized with the DNA fragments. Enrichment for the targeted sequences is achieved after the capture of the hybrids on streptavidin-coated beads and magnetic separation from irrelevant sequences. PCR-adapter-compatible primers (*blue*) (optionally barcoded—*light-green*) amplify the captured fragments. The ligation of the protein-conjugated sequencing-adapters is followed as normal afterward

2.1 Enzymes

1. Platinum SYBR Green qPCR SuperMix-UDG and Platinum Taq DNA Polymerase (ThermoFisher Scientific, Waltham, MA, USA).
2. NEBNext End Repair module, NEBNext dA tailing module, NEBNext Blunt/TA Ligase, and LongAmp Taq 2× Master Mix (New England Biolabs, Ipswich, MA, USA).

2.2 Additional Kits

1. SeqCap Hybridization and Wash Kits (Roche Diagnostics, Indianapolis, IN, USA).
2. QIAquick Gel Extraction Kit (Qiagen, Valencia, CA, USA).
3. Quant-iT PicoGreen dsDNA (ThermoFisher Scientific).

2.3 Other Reagents and Consumables

1. g-TUBEs (Covaris, Woburn, MA, USA).
2. Agencourt AMPure XP PCR Purification beads (Beckman Coulter, Carlsbad, CA, USA).
3. Dynabeads M-270 Streptavidin (ThermoFisher Scientific).

4. Human Cot-1 DNA.
5. NaCl and Tris-HCl, EDTA for making the hybridization buffer (10 mM Tris-HCl, 1 mM EDTA pH 7.5–8.0, 50 mM NaCl).

3 Methods

After extended optimization of the baits' length and the shearing size of the genomic DNA, we concluded that hybridizing ~500 bp long baits with ~5000 bp long DNA fragments would give the best enrichment results [12]. Thus, we have PCR-generated 3 baits, of average length 460 bp, to target a 1317 bp region of Phage lambda genome (NC_001416.1, nt. 41,053–42,370). For targeting the rRNA operons of *E. coli* strains O157:H7 (NC_002695) and K12/W3110 (NC_007779.1) we have generated 9 baits, of average length 597 bp, covering the 5258 bp-long *rrnH* operon of strain O157:H7 (nt. 227,102–232,360) (*see Note 1*).

The described protocol below has been updated to conform to the latest MinION sequencing and barcoding kits and reagents:

3.1 Preparation of the Baits

1. Purify the PCR products to be used as baits or gel-extract them in case of multiple bands appear in the agarose gel.
2. Normalize the concentrations of the PCR products using Quant-iT PicoGreen (*see Note 2*).
3. Mix them in equimolar concentrations and dilute ~4 µg in 85 µL of TE buffer.
4. Blunt-end-repair the mix by adding 85 µL of baits mix, 10 µL of reaction buffer and 5 µL of enzyme mix and incubating at 20 °C for 30 min.
5. Purify with 1.8× Agencourt AMPure XP beads, elute in 30 µL of TE buffer.
6. Construct the biotin adapters as described in [11], using the hybridization buffer (10 mM Tris-HCl, pH 7.5–8.0, 1 mM EDTA, 50 mM NaCl) to rehydrate the biotin conjugated oligos. Denature the oligos at 95 °C for 5 min and gradually anneal them ramping down to 20 °C for 75 min. Keep them on ice until use.
7. Ligate the biotin adapters to the end-repaired PCR-products by adding 30 µL of end-repaired baits mix, 20 µL of biotin adapters, and 50 µL of 2× Blunt/TA Ligase mastermix, incubate at room temperature for 20 min.
8. Purify with 1.8× Agencourt AMPure XP beads, elute in 16 µL of TE buffer.
9. Quantify the resulting baits with PicoGreen and keep them on ice.

3.2 Preparation of the Genomic DNA

1. Shear the genomic DNA (2 μg of DNA in 85 μL TE) in 5000–6000 bp fragments using g-Tubes.
2. Blunt-end repair by adding: 85 μL of sheared DNA, 10 μL of reaction buffer, and 5 μL of enzyme mix and incubating at 20 °C for 30 min.
3. Purify with 1 \times Agencourt AMPure XP beads, elute in 25 μL of TE buffer.
4. Perform dA-tailing by adding: 25 μL of end-repaired genomic DNA, 3 μL of reaction buffer and 2 μL of enzyme.
5. Ligate the PCR adapters (PCA—included in the Genomic DNA Sequencing kit) by adding: 30 μL of dA-tailed Genomic DNA, 20 μL of MinION PCR adapters, and 50 μL of 2 \times Blunt/TA Ligase mastermix, incubate at room temperature for 20 min (*see Note 3*).
6. Purify with 1 \times Agencourt AMPure XP beads, elute in 16 μL of TE buffer.
7. Quantify the resulting PCR-adaptor-ligated Genomic DNA with PicoGreen (use 1 μL) and keep it in ice.

3.3 Hybridization and Capture

1. Follow the protocol of SeqCap Hybridization and Wash Kit after modifying the initial hybridization mixture: add 500 ng of genomic DNA (diluted in 15 μL TE) (final volume/2) μL of 2 \times Hybridization buffer and (final volume/10) μL Hybridization component A.
2. Denature the hybridization mix at 95 °C for 5 min, and then add immediately 8 μL of baits (*see Note 4*).
3. Follow the *SeqCap Hybridization and Wash Kit* protocol up to the final washing step of streptavidin beads.
4. Resuspend the pelleted beads in 48 μL of PCR-grade water.
5. Amplify the captured fragments using the provided Primer Mix (PRM—included in the Genomic DNA Sequencing kit) (*see Note 5*). Add 50 μL of LongAmp Taq 2 \times mastermix (M0287S, New England BioLabs, Hitchin, UK), 2 μL of MinION primers (barcoded if necessary) and the resuspended beads. Run PCR program as follows: initial denaturation 95 °C for 3 min, 18 amplification cycles of 95 °C for 15 s, 62 °C for 15 s, 65 °C for 4 min, final elongation 65 °C for 8 min and 4 °C hold (*see Note 6*).
6. Remove the streptavidin beads and the baits with a magnetic rack and purify the product with 1 \times AMPure XP beads.
7. Quantify with Picogreen and dilute 1 μg of PCR product in 80 μL of PCR-grade water or TE.

3.4 MinION Sequencing-Adapters Ligation and Library Preparation

The PCR-amplified captured library can now be used as template for the MinION library preparation (*see Note 7*).

1. Add the provided internal control (CS-DNA).
2. Follow the MinION genomic DNA library preparation protocol for the end-repair, dA-tailing, and the ligation of sequencing-adapters.
3. Condition and load the library to the sequencer, reloading library in standard intervals.
4. Bioinformatics Analysis can be performed using MinION dedicated tools (*see Note 8*).

4 Notes

1. Alternative method for generating the baits: For targets exceeding ~10 Kb in total, multiple PCR reactions can be labor to be developed and optimized, especially in the case of complex templates. Alternatively, the baits can be purchased as synthesized dsDNA fragments of ~600 bp each, which can either be biotinylated or can be ligated to the biotin adapters as described here and in [11].
2. Nanodrop or QuBit fluorimeter can be used alternatively.
3. In the case of multiplexed libraries, the Barcode Adapter (BCA—included in the PCR Barcoding Kit) should be used instead.
4. The copy number of the baits should be approximately 1000× the copies of the target sequence.
5. In the case of multiplexed libraries, the Barcode Primers (BC1–12, included in the PCR Barcoding Kit) should be used instead (*see Table 1*).
6. It is not recommended to keep the captured library in ice for long or overnight. If possible continue directly to MinION library preparation.
7. Quality Control of the Enrichment: The capture of the target sequences can be assessed with SYBR® Green qPCR before the ligation of the MinION sequencing adapters: Design primers complimentary to the targeted and several [2, 3] untargeted regions. Perform all the qPCR reactions in parallel, using as a template both the captured and the initial genomic DNA. Increment of the relative ΔC_q after the capture is indicative of successful enrichment.
8. Bioinformatics and Tools for MinION Data Analysis. Base calling is now performed via EPI2ME operating system (provided by Metrichor company). EPI2ME operating system is coupled with MinION and enables immediate and ongoing

Table 1
Barcode sequences used in the primers of the “PCR barcoding kit”

Primer	5'-Sequence-3'
BC1	GGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCT
BC2	GGTGCTGTCGATTCGGTTTGTAGTCGTCTGTTAACCT
BC3	GGTGCTGGAGTCTTGTGTCCCAGTTACCAGGTTAACCT
BC4	GGTGCTGTTTCGGATTCTATCGTGTTCCTATTAACCT
BC5	GGTGCTGCTTGTCCAGGGTTTGTGTAACCTTTAACCT
BC6	GGTGCTGTTCTCGCAAAGGCAGAAAGTAGTCTTAACCT
BC7	GGTGCTGGTGTACCCTGGGAATGAATCCTTTAACCT
BC8	GGTGCTGTTTCAGGGAACAAACCAAGTTACGTTAACCT
BC9	GGTGCTGAACTAGGCACAGCGAGTCTTGGTTTTAACCT
BC10	GGTGCTGAAGCGTTGAAACCTTTGTCTCTCTTAACCT
BC11	GGTGCTGGTTTCATCTATCGGAGGGAATGGATTAACCT
BC12	GGTGCTGCAGGTAGAAAGAAGCAGAATCGGATTAACCT

interrogation of the data. However, an open source alternative has been recently described [13]. A useful tool for converting the *.fast5* reads to *.fasta/q* and accessing run statistics is *poRe* package [5] for R programming language. LAST aligner [14] can be used for mapping the MinION reads to the respective references. As an end-to-end optimized aligner is very capable of handling longer reads. Alignments produced by LAST are in *.maf* file format and need conversion (to *.bam* files). This can be done with the widely available “maf-convert” Python script. Downstream analysis for variation calling etc. is the same as in other platforms. Dedicated tools are available though (<https://github.com/mitenjain/nanopore>). Nanocorrect (<https://github.com/jts/nanocorrect>), nanoCORR (<https://github.com/jgurtowski/nanocorr>), nanopolish (<https://github.com/jts/nanopolish>) and proovread (<https://github.com/BioInf-Wuerzburg/proovread>) are recently released tools that are designed to improve the quality of MinION reads with or without the assistance of data derived from other platforms. Barcoded libraries can be demultiplexed automatically through Metrichor applications. Several tools can be used instead, like “FASTQ/A Barcode splitter” from “fastx toolkit” (http://hannonlab.cshl.edu/fastx_toolkit), “fastq-multx” from “ea-utils” (<http://code.google.com/p/ea-utils>) and others, given the MinION barcode sequences (see Table 1).

References

1. Kasianowicz JJ, Brandin E, Branton D, Deamer DW (1996) Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci U S A* 93(24):13770–13773
2. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich SB, Krstic PS, Lindsay S, Ling XS, Mastrangelo CH, Meller A, Oliver JS, Pershin YV, Ramsey JM, Riehn R, Soni GV, Tabard-Cossa V, Wanunu M, Wiggin M, Schloss JA (2008) The potential and challenges of nanopore sequencing. *Nat Biotech* 26(10):1146–1153. <https://doi.org/10.1038/nbt.1495>
3. Stoddart D, Heron AJ, Mikhailova E, Maglia G, Bayley H (2009) Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proc Natl Acad Sci U S A* 106(19):7702–7707. <https://doi.org/10.1073/pnas.0901054106>
4. Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, Akeson M (2012) Automated forward and reverse ratcheting of DNA in a nanopore at 5-a precision. *Nat Biotech* 30(4):344–348. <https://doi.org/10.1038/nbt.2147>
5. Watson M, Thomson M, Risse J, Talbot R, Santoyo-Lopez J, Gharbi K, Blaxter M (2014) poRe: an R package for the visualization and analysis of nanopore sequencing data. *Bioinformatics* 31(1):114–115. <https://doi.org/10.1093/bioinformatics/btu590>
6. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M (2015) Improved data analysis for the MinION nanopore sequencer. *Nat Meth* 12(4):351–356. <https://doi.org/10.1038/nmeth.3290>
7. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR (2007) Genome-wide in situ exon capture for selective resequencing. *Nat Genet* 39(12):1522–1527. <https://doi.org/10.1038/ng.2007.42>
8. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotech* 27(2):182–189. <https://doi.org/10.1038/nbt.1523>
9. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ (2010) Target-enrichment strategies for next-generation sequencing. *Nat Meth* 7(2):111–118. <https://doi.org/10.1038/nmeth.1419>
10. Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Schmitz R, Doronichev VB, Golovanova LV, de la Rasilla M, Fordea J, Rosas A, Paabo S (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* 325(5938):318–321. <https://doi.org/10.1126/science.1174462>
11. Maricic T, Whitten M, Paabo S (2010) Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One* 5(11):e14004. <https://doi.org/10.1371/journal.pone.0014004>
12. Karamitros T, Magiorkinis G (2015) A novel method for the multiplexed target enrichment of MinION next generation sequencing libraries using PCR-generated baits. *Nucl Acid Res* 43(22):e152. <https://doi.org/10.1093/nar/gkv773>
13. David M, Dursi LJ, Yao D, Boutros PC, Simpson JT (2016) Nanocall: an open source basecaller for oxford nanopore sequencing data. *bioRxiv* 33(1). <https://doi.org/10.1101/046086>
14. Frith MC, Hamada M, Horton P (2010) Parameters for accurate genome alignment. *BMC Bioinform* 11:80. <https://doi.org/10.1186/1471-2105-11-80>

Hi-Plex for Simple, Accurate, and Cost-Effective Amplicon-based Targeted DNA Sequencing

Bernard J. Pope, Fleur Hammet, Tu Nguyen-Dumont, and Daniel J. Park

Abstract

Hi-Plex is a suite of methods to enable simple, accurate, and cost-effective highly multiplex PCR-based targeted sequencing (Nguyen-Dumont et al., *Biotechniques* 58:33–36, 2015). At its core is the principle of using gene-specific primers (GSPs) to “seed” (or target) the reaction and universal primers to “drive” the majority of the reaction. In this manner, effects on amplification efficiencies across the target amplicons can, to a large extent, be restricted to early seeding cycles. Product sizes are defined within a relatively narrow range to enable high-specificity size selection, replication uniformity across target sites (including in the context of fragmented input DNA such as that derived from fixed tumor specimens (Nguyen-Dumont et al., *Biotechniques* 55:69–74, 2013; Nguyen-Dumont et al., *Anal Biochem* 470:48–51, 2015), and application of high-specificity genetic variant calling algorithms (Pope et al., *Source Code Biol Med* 9:3, 2014; Park et al., *BMC Bioinformatics* 17:165, 2016). Hi-Plex offers a streamlined workflow that is suitable for testing large numbers of specimens without the need for automation.

Key words Amplicon sequencing, Targeted sequencing, Massively parallel sequencing, Next-generation sequencing, Multiplex PCR, Hi-Plex, Sequence screening, Mutation screening

1 Introduction

Hi-Plex (Fig. 1) was conceived for highly accurate, streamlined, and economical sequence-screening of large specimen numbers. It is compatible with the highest fidelity PCR polymerases available today and permits the order of 1000+ barcoded amplicons to be produced in a single reaction chamber without automation or phase separation (we have demonstrated a range of 16–1003 amplicons per reaction vessel) [1–3]. Products resulting from hundreds (potentially thousands) of specimen reactions can then be combined for size selection on a single lane of an agarose gel [6–8]. Following simple gel excision and DNA purification, the library is ready for quantitation and DNA sequencing.

Alternative approaches for amplicon-based targeted sequence screening, such as Ampliseq, TruSeq Amplicon, Fluidigm, or

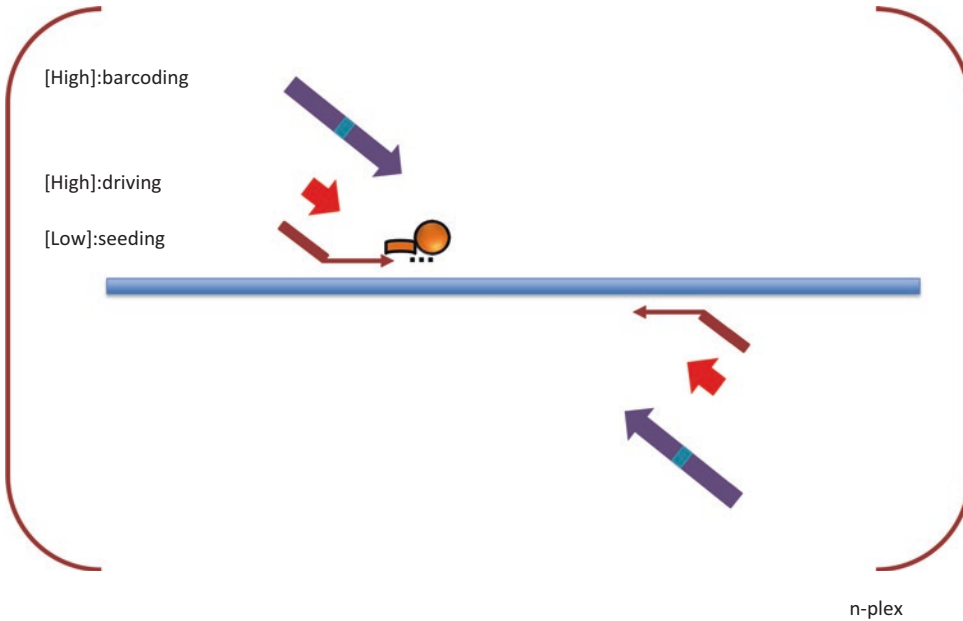


Fig. 1 Schematic illustration of the Hi-Plex PCR mechanisms. In a single reaction, 5' “heeled” GSPs representing all targeted amplicons (n amplicons) at low individual concentrations seed the reaction and relatively high concentrations of abridged adapter primers (universal primers) (shown in *red*) drive PCR. Full-length adapter primers (barcoding adapter primers) are included in the PCR only during terminal cycles

Haloplex, are, variously, expensive, require complicated enzyme reaction series, confer primer design constraints, involve cumbersome parallel processing workflows, require expensive, dedicated machinery or only allow the use of chemistries with relatively low accuracy.

To seed the reaction, Hi-Plex uses a low concentration of gene-specific primers (GSPs) (on an individual primer basis—the aggregate concentration is relatively high) that have been engineered to contain universal 5' “heel” sequences. Higher concentration, relatively short universal primers drive the majority of amplification and restrict GSP efficiency biases to very early PCR cycles. Barcoding adapter primers are added to the reaction during latter cycles to minimize off-target priming effects. GSPs are designed so that target regions are “tiled” and resulting amplicons fall within a narrow size range. Besides helping with the uniformity of representation across the target amplicons, this allows us to use focused size selection as a stringent purification step. It also allows us, optionally, to define amplicon sizes so that we can achieve complete overlap of the queried sequence for both reads of a read-pair following paired-end sequencing. In turn, this allows the application of high-accuracy genetic variant calling algorithms that require both reads of a pair to agree for a given position for those reads to contribute to making “calls” (Fig. 2). Importantly, the size selection step needs only to be performed once (on a single gel lane) for the screening of hundreds of specimens

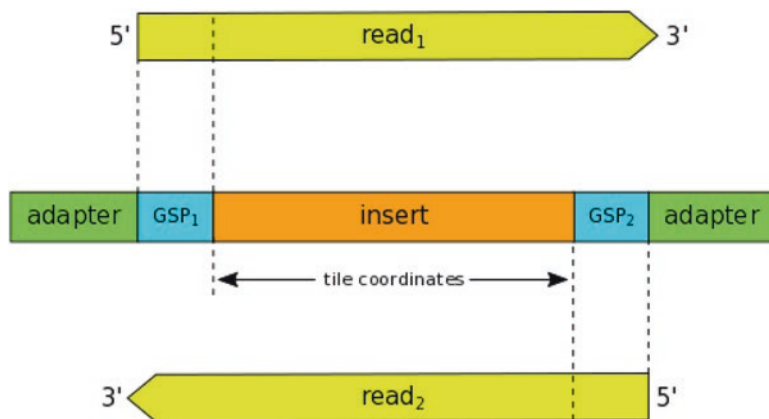


Fig. 2 Hi-Plex library structure and (optional) overlapping reads to work with ROVER and UNDR-ROVER variant-calling software. The *centre rectangle* represents the target insert DNA sequence flanked by gene-specific primer (GSP) sites (*blue*) and adapter sequences (*green*). The two reads of a pair are shown in *yellow*. The 5' end of each read starts with its corresponding gene-specific primer sequence. The insert size is chosen so that both reads overlap the target insert sequence completely. The 3' ends of reads may extend into the adapter sequence depending on the read length and the presence/absence of insertions/deletions in the template DNA. The diagram is not to scale. Typically, the insert sequence will be significantly longer than the primer sequences

because the Hi-Plex reaction products are barcoded and can be pooled together prior to size separation-based purification.

The following presents an example 688-plex library protocol for sequencing using the Illumina MiSeq instrument (*see Note 1*).

2 Materials

2.1 Template DNAs

Template DNAs should be prepared using a high quality DNA extraction system and quantified as accurately as possible by Picogreen assay (or equivalent) and calibrated instrumentation (*see Note 2*).

2.2 General Labware and Instrumentation

1. PCR machine (PCR plate must be compatible with the thermocycler).
2. Qubit (ThermoFisher Scientific, Waltham, MA, USA).
3. BioAnalyzer (Agilent Technologies, Santa Clara, CA, USA).
4. 96-well PCR plate.
5. Narrow stem transfer pipette (ThermoFisher Scientific).
6. Eppendorf low binding tubes (Eppendorf, Hamburg, Germany).
7. MicroAmp Clear adhesive film (ThermoFisher Scientific).

2.3 Solutions

1. Low TE: 10 mM Tris–HCl, 0.1 mM EDTA, pH 8.0.
2. Molecular biology grade water.
3. 20 mM dNTPs: 20 mM dATP, 20 mM dTTP, 20 mM dCTP, 20 mM dGTP.
4. Phusion Hot Start High-Fidelity DNA Polymerase (ThermoFisher Scientific)

2.4 PCR Primers

All the primer sequences are shown in the 5' to 3' orientation. All the primers are suspended in low TE and can be ordered at standard desalting grade with the exception of the custom sequencing primers, which should be HPLC purified (or equivalent).

2.4.1 GSPs

Combine 5 µL of each individual GSP at 200 µM into a single GSP pool vessel. We take this to represent an aggregate GSP pool concentration of 200 µM. Create a working stock of 50 µM aggregate GSP pool by diluting an aliquot with molecular grade water.

Example forward GSP (lower case shows universal heel sequence, upper case shows gene-specific sequence):
ctctctatgggcagtcggtgattTGGTAATAATTTTAGGACACTG-TAGTTCCTG

Example reverse GSP (lower case shows universal heel sequence, upper case shows gene-specific sequence):
ctgcgtgtctccgactcagGAACCAATATACAGGACAATGAGTC-TACA

2.4.2 Universal Primers and Custom Sequencing Primers

1. Universal primer F1: ctctctatgggcagtcggtgatt
2. Universal primer R1: ctgcgtgtctccgactcag
3. Read 1 primer: ccatctcatccctgcgtgtctccgactcag
4. Read 2 primer: ctccgctttctctctatgggcagtcggtgatt
5. i7 read primer: aatcaccgactgccatagaggaagcggag
6. Dual indexing barcoding adapter primers (lower case sequences correspond to Read 1 primer or Read 2 primer sites. The eight bases immediately preceding Read 1 or Read 2 sites are “barcodes”):

N501_TSIT_A	AATGATACGGCGACCACCGAGATCTACACTAGATCGC ccatctcatccctgcgtgtctccgactcag
N502_TSIT_A	AATGATACGGCGACCACCGAGATCTACACCTCTCTAT ccatctcatccctgcgtgtctccgactcag
N503_TSIT_A	AATGATACGGCGACCACCGAGATCTACACTATCCTCT ccatctcatccctgcgtgtctccgactcag
N504_TSIT_A	AATGATACGGCGACCACCGAGATCTACACAGAGTAGA ccatctcatccctgcgtgtctccgactcag

(continued)

N505_TSIT_A	AATGATACGGCGACCACCGAGATCTACACGTAAGGAG ccatctcatccctgcgtgtctccgactcag
N506_TSIT_A	AATGATACGGCGACCACCGAGATCTACACACTGCATA ccatctcatccctgcgtgtctccgactcag
N507_TSIT_A	AATGATACGGCGACCACCGAGATCTACACAAGGAGTA ccatctcatccctgcgtgtctccgactcag
N508_TSIT_A	AATGATACGGCGACCACCGAGATCTACACCTAAGCCT ccatctcatccctgcgtgtctccgactcag
N509_TSIT_A	AATGATACGGCGACCACCGAGATCTACACTCGCTAGA ccatctcatccctgcgtgtctccgactcag
N510_TSIT_A	AATGATACGGCGACCACCGAGATCTACACCTATCTCT ccatctcatccctgcgtgtctccgactcag
N511_TSIT_A	AATGATACGGCGACCACCGAGATCTACACCTTTATC ccatctcatccctgcgtgtctccgactcag
N512_TSIT_A	AATGATACGGCGACCACCGAGATCTACACTAGAAGAG ccatctcatccctgcgtgtctccgactcag
N513_TSIT_A	AATGATACGGCGACCACCGAGATCTACACGGAGGTAA ccatctcatccctgcgtgtctccgactcag
N514_TSIT_A	AATGATACGGCGACCACCGAGATCTACACCATAACTG ccatctcatccctgcgtgtctccgactcag
N515_TSIT_A	AATGATACGGCGACCACCGAGATCTACACAGTAAAGG ccatctcatccctgcgtgtctccgactcag
N516_TSIT_A	AATGATACGGCGACCACCGAGATCTACACGCCTCTAA ccatctcatccctgcgtgtctccgactcag
N517_TSIT_A	AATGATACGGCGACCACCGAGATCTACACGATCGCTA ccatctcatccctgcgtgtctccgactcag
N518_TSIT_A	AATGATACGGCGACCACCGAGATCTACACCTCTATCT ccatctcatccctgcgtgtctccgactcag
N519_TSIT_A	AATGATACGGCGACCACCGAGATCTACACTCCTCTTA ccatctcatccctgcgtgtctccgactcag
N520_TSIT_A	AATGATACGGCGACCACCGAGATCTACACAGTAGAAG ccatctcatccctgcgtgtctccgactcag
N521_TSIT_A	AATGATACGGCGACCACCGAGATCTACACAAGGAGGT ccatctcatccctgcgtgtctccgactcag
N522_TSIT_A	AATGATACGGCGACCACCGAGATCTACACTGCATAAC ccatctcatccctgcgtgtctccgactcag
N523_TSIT_A	AATGATACGGCGACCACCGAGATCTACACGGAGTAAA ccatctcatccctgcgtgtctccgactcag
N524_TSIT_A	AATGATACGGCGACCACCGAGATCTACACAAGCCTCT ccatctcatccctgcgtgtctccgactcag

(continued)

N701_TSIT_P	CAAGCAGAAGACGGCATAACGAGATTTCGCCTT ctccgctttcctctctatgggcagtcggtgat
N702_TSIT_P	CAAGCAGAAGACGGCATAACGAGATCTAGTACG ctccgctttcctctctatgggcagtcggtgat
N703_TSIT_P	CAAGCAGAAGACGGCATAACGAGATTTCTGCCT ctccgctttcctctctatgggcagtcggtgat
N704_TSIT_P	CAAGCAGAAGACGGCATAACGAGATGCTCAGGA ctccgctttcctctctatgggcagtcggtgat
N705_TSIT_P	CAAGCAGAAGACGGCATAACGAGATAGGAGTCC ctccgctttcctctctatgggcagtcggtgat
N706_TSIT_P	CAAGCAGAAGACGGCATAACGAGATCATGCCTA ctccgctttcctctctatgggcagtcggtgat
N707_TSIT_P	CAAGCAGAAGACGGCATAACGAGATGTAGAGAG ctccgctttcctctctatgggcagtcggtgat
N708_TSIT_P	CAAGCAGAAGACGGCATAACGAGATCCTCTCTG ctccgctttcctctctatgggcagtcggtgat
N709_TSIT_P	CAAGCAGAAGACGGCATAACGAGATAGCGTAGC ctccgctttcctctctatgggcagtcggtgat
N710_TSIT_P	CAAGCAGAAGACGGCATAACGAGATCAGCCTCG ctccgctttcctctctatgggcagtcggtgat
N711_TSIT_P	CAAGCAGAAGACGGCATAACGAGATTGCCTCTT ctccgctttcctctctatgggcagtcggtgat
N712_TSIT_P	CAAGCAGAAGACGGCATAACGAGATTCCTCTAC ctccgctttcctctctatgggcagtcggtgat
N713_TSIT_P	CAAGCAGAAGACGGCATAACGAGATCTTATCGC ctccgctttcctctctatgggcagtcggtgat
N714_TSIT_P	CAAGCAGAAGACGGCATAACGAGATTACGCTAG ctccgctttcctctctatgggcagtcggtgat
N715_TSIT_P	CAAGCAGAAGACGGCATAACGAGATGCCTTTCT ctccgctttcctctctatgggcagtcggtgat
N716_TSIT_P	CAAGCAGAAGACGGCATAACGAGATAGGAGCTC ctccgctttcctctctatgggcagtcggtgat
N717_TSIT_P	CAAGCAGAAGACGGCATAACGAGATGTCCAGGA ctccgctttcctctctatgggcagtcggtgat
N718_TSIT_P	CAAGCAGAAGACGGCATAACGAGATCCTACATG ctccgctttcctctctatgggcagtcggtgat
N719_TSIT_P	CAAGCAGAAGACGGCATAACGAGATAGAGGTAG ctccgctttcctctctatgggcagtcggtgat
N720_TSIT_P	CAAGCAGAAGACGGCATAACGAGATTCTGCCTC ctccgctttcctctctatgggcagtcggtgat
N721_TSIT_P	CAAGCAGAAGACGGCATAACGAGATTAGCAGCG ctccgctttcctctctatgggcagtcggtgat

(continued)

N722_ TSIT_P	CAAGCAGAAGACGGCATAACGAGATCTCGCAGC ctccgctttcctctctatgggcagtcggtgat
N723_ TSIT_P	CAAGCAGAAGACGGCATAACGAGATTCTTTGCC ctccgctttcctctctatgggcagtcggtgat
N724_ TSIT_P	CAAGCAGAAGACGGCATAACGAGATCTACTCCT ctccgctttcctctctatgggcagtcggtgat
N725_ TSIT_P	CAAGCAGAAGACGGCATAACGAGATGCCTTATC ctccgctttcctctctatgggcagtcggtgat
N726_ TSIT_P	CAAGCAGAAGACGGCATAACGAGATAGTACGCT ctccgctttcctctctatgggcagtcggtgat
N727_ TSIT_P	CAAGCAGAAGACGGCATAACGAGATCTGCCTTT ctccgctttcctctctatgggcagtcggtgat
N728_ TSIT_P	CAAGCAGAAGACGGCATAACGAGATTCAGGAGC ctccgctttcctctctatgggcagtcggtgat
N729_ TSIT_P	CAAGCAGAAGACGGCATAACGAGATGAGTCCAG ctccgctttcctctctatgggcagtcggtgat
N730_ TSIT_P	CAAGCAGAAGACGGCATAACGAGATTGCCTACA ctccgctttcctctctatgggcagtcggtgat
N731_ TSIT_P	CAAGCAGAAGACGGCATAACGAGATAGAGAGGT ctccgctttcctctctatgggcagtcggtgat
N732_ TSIT_P	CAAGCAGAAGACGGCATAACGAGATTCTCTGCC ctccgctttcctctctatgggcagtcggtgat
N733_ TSIT_P	CAAGCAGAAGACGGCATAACGAGATCGTAGCAG ctccgctttcctctctatgggcagtcggtgat
N734_ TSIT_P	CAAGCAGAAGACGGCATAACGAGATGCCTCGCA ctccgctttcctctctatgggcagtcggtgat
N735_ TSIT_P	CAAGCAGAAGACGGCATAACGAGATCCTCTTTG ctccgctttcctctctatgggcagtcggtgat
N736_ TSIT_P	CAAGCAGAAGACGGCATAACGAGATCTTACTC ctccgctttcctctctatgggcagtcggtgat

2.5 Agarose Gel Electrophoresis

1. UltraPure Agarose1000 (ThermoFisher Scientific).
2. 1×TBE (prepared from 10× ultrapure TBE).
3. 10 mg/mL ethidium bromide.
4. 50 bp-incremented DNA ladder.
5. Preparative gel comb.
6. Gel Pilot 5× Loading Dye (Qiagen, Hilden, Germany).
7. DNA Gel Extraction Kit: QiaEX II Gel Extraction Kit (Qiagen)

2.6 DNA Quantitation and Quality Assessment

1. Qubit ds DNA High Sensitivity Assay Kit (ThermoFisher Scientific).
2. BioAnalyzer reagents: Agilent High Sensitivity DNA chip kit (Agilent Technologies).
3. 300 bp MiSeq reagent kit v2 (Illumina, San Diego, CA, USA) (*see Note 3*).
4. HiSeq Install Accessories box (Illumina).

3 Methods

3.1 Primer Design

Refer to Materials for example sequences. The GSPs, each comprised of a 3' gene-specific portion and a 5' universal portion, should be designed so that the resulting amplicons vary little in size (typically, the range of size difference would be about 30 bp). The authors frequently design the library “inserts” (i.e., the sequence to be queried, excluding the gene-specific primer portion) to be close to 100 bp (*see Note 4*). The gene-specific primers should also be designed to minimize variability in melting temperature and propensity for primer-dimer formation (*see Note 5*). The target melting temperature should be set to consider reaction stringency and design constraints (*see Note 6*). The sequence introduced by the 5' universal portions of the GSPs are used to template universal primer-driven amplification for the majority of Hi-Plex amplification and subsequent barcoding adapter primer extension. They are also used to template sequencing reactions, along with additional sequences introduced by the barcoding adapter primers (*see Note 7*).

3.2 Hi-Plex PCR

1. Prepare a template document to map the pattern of dual index barcoding adapter primer-pairs to identify each DNA specimen in the PCR plate (*see Note 8*).
2. Prepare a 10 μ M (each barcoding adapter primer should be at 10 μ M) working stock plate of barcoding adapter primer-pairs arrayed in the designated pattern. To avoid the risk of PCR contamination, re-use of working stock barcoding adapter primer plates should be minimized. Higher concentration stock plates can be stored frozen in a no-copy facility for ease of working stock generation.
3. Importantly, the following PCR set-up steps should be conducted *on ice* so that, when ready, the PCR plate can be transferred immediately from ice to the PCR machine block preheated (paused) at 98 °C (*see Note 9*). Prepare a PCR master mix *on ice* by pipetting up and down and without vortexing. The following example is typical for applying Hi-Plex to 96 samples in a 96-well plate (*see Note 10*):

PCR master mix:

Reagent	Conc.	μL for single well	[Final]	μL for 96-well plate ($\times 110$)
Molecular grade water		12.5		1375
5 \times HF Phusion buffer	5 \times	5	1 \times	550
MgCl ₂	50 mM	0.5	2.5 mM	55
Phusion hot start HF	2 U/ μL	0.5	1 U	55
dNTPs	20 mM	0.5	400 μM	55
GSPs (aggregate conc.)	50 μM	0.75	1.5 μM	82.5
Universal primers F1/R1 (aggregate conc.)	100 μM	0.25	1 μM	27.5
	Total	20		2200

- Add 20 μL of ice-cold PCR master mix to each well of a thin-walled PCR plate on ice. For each specimen, add 2.5 μL of ice-cold 10 ng/ μL DNA to its designated well and pipette up and down to mix (*see Note 11*). Seal the plate completely with film that is compatible with a heated lid thermocycler.
- Apply the following thermocycling regimen, using a calibrated PCR machine that is compatible with the PCR plate molding and with the *heated lid function turned on* (*see Note 12*). The machine should be paused at 98 °C so that the PCR plate can be transferred directly from ice to this thermal denaturing condition prior to thermocycling:

STEP	Temp.	Time	# cycles
Enzyme activation	98 °C	1 min	1
Target amplification	98 °C	30 s	16
	58 °C	1 min	
	62 °C	1 min	
	65 °C	1 min	
	72 °C	1 min	
HOLD	70 °C		

- Immediately upon reaching the 70 °C HOLD stage of thermocycling, add 2.5 μL of 10 μM barcoding adapter primer-pairs (aggregate concentration of forward and reverse primers) to the designated wells of the PCR plate. Use a multichannel pipette to aspirate barcoding adapter primer-pairs from the adapter array plate prepared earlier. Remove the adhesive film from the PCR plate (keeping the plate in the thermocycler at 70 °C) and add the barcoding adapter primers to their designated wells. Reseal the PCR plate with fresh film.

7. Immediately proceed from the 70 °C HOLD to continue thermocycling using the following parameters (*see Note 13*):

Full length adapter extension cycling	98 °C	30 s	2–4
	68 °C	2 min	
	72 °C	1 min	
Extension	68 °C	20 min	1
HOLD	4 °C		

3.3 Hi-Plex Size-Selection

1. Generate a multi-specimen pooled library by collecting 8 µL of product for each individual specimen together into a single tube. The extent of specimen multiplexing depends on the number to be analyzed per sequencing lane.
2. Prepare a 1×TBE running buffer and a 1.75% agarose (w/v) (1×TBE) gel containing 500 ng/µL ethidium bromide. Use a gel comb with wide teeth (e.g., 26 mm × 1.5 mm) for the library and narrower teeth for size markers. Customized, wide teeth can be achieved by taping multiple smaller teeth.
3. Load the gel using 10 µL of 40 ng/µL 50 bp-incremented DNA ladder in a narrow well and 70 µL of pooled library (i.e., 70 µL pooled library plus 14 µL of 5× GelPilot loading dye) in a wide well (*see Note 14*). Conduct gel electrophoresis at 100 V for 90 min.
4. Assess the library gel under 302 nm UV light and use a clean scalpel blade to excise the library band (in our example application, the product size is ~270–280 bp). Photograph the gel before and after excision (*see Note 15*). Figure 3a depicts a typical gel profile prior to excision.
5. Purify the library using a gel extraction kit (any library gel slices that are excess to requirements can be stored at 4 °C). Typically, we use the QiaEX II system. Estimate the volume of excised agarose by weight using a balance that is capable of accurately measuring down to 10 mg. Assuming 1 mL to approximate 1 g, add three volumes of QX1 buffer to the tube containing the excised library (e.g., 250 mg of gel slice requires 750 mL of QX1 solution). Vortex the QiaEX II bead slurry thoroughly and add 10 µL to the gel and QX1 mixture. Continue the purification according to the manufacturer's instructions, except that the library should be eluted using 20 µL of low TE. Aim to aspirate 15 µL of supernatant without disturbing the bead pellet. If multiple lanes have been run and products purified for a given pooled library (e.g., in the case of low library yield), the resulting eluates can be combined at this point). For example, the eluates from the two corresponding pooled library lanes could be combined to result in 30 µL of purified pooled library.

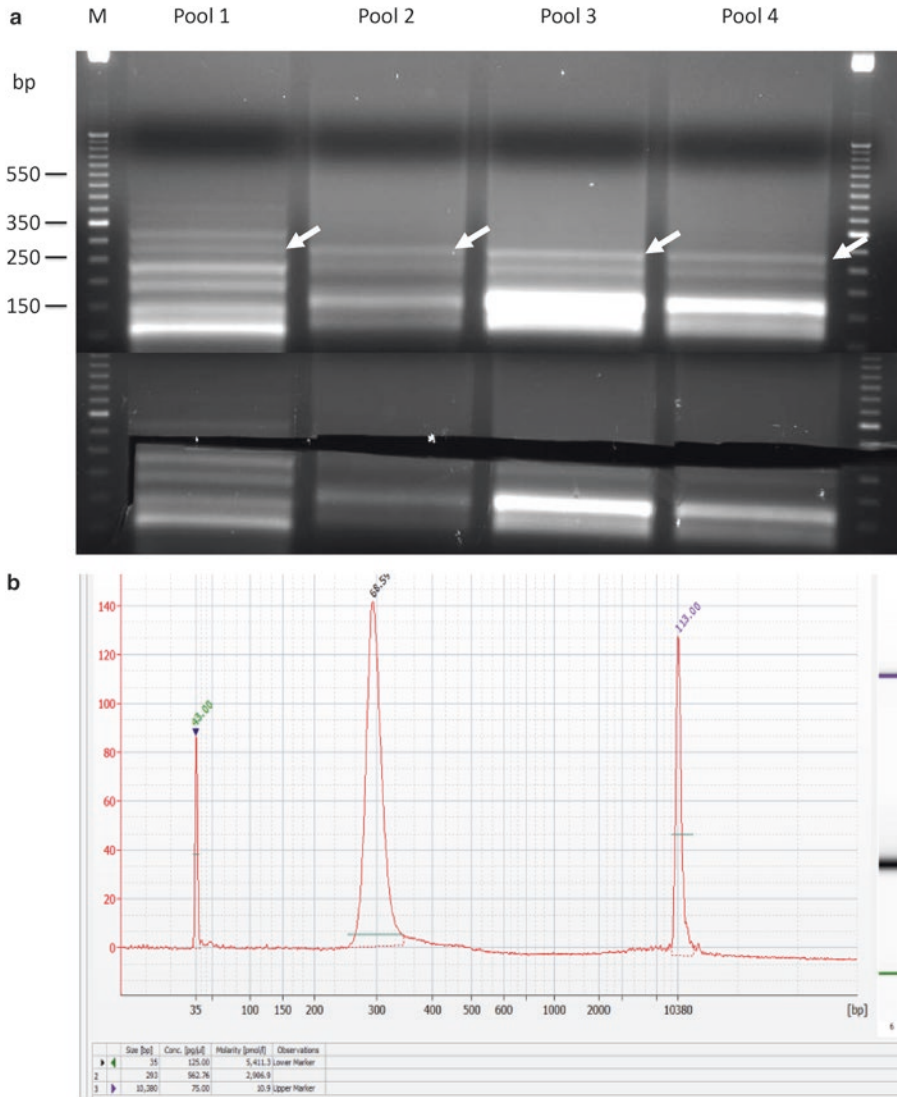


Fig. 3 (a) Representative agarose gel images illustrating the excision of a variety of different multi-specimen Hi-Plex libraries for different targets and levels of amplicon-plexity. Arrows indicate the libraries. “M” indicates a size standard marker. **(b)** A representative BioAnalyser profile for a multi-specimen Hi-Plex library produced using the current example protocol (688 amplicons for each specimen). The *y*-axis numbers relate to fluorescence units (FUs). The outer peaks represent the reference markers and the central peak represents the multi-specimen library

6. Measure the molarity of the library using the Qubit High sensitivity DNA kit according to the manufacturer’s instructions (*see Note 16*). Use 5 μ L of library to quantitate by Qubit. Ultimately, we are aiming to adjust the library to 450 $\text{pg}/\mu\text{L}$ which equates to the approx. 2.3 nM that our current example protocol requires for denaturation prior to loading on the MiSeq (*see Note 17*).

7. Analyze the library by Agilent BioAnalyzer. Ensure that the library is in the range of 100–500 pg/ μ L based on Qubit measurement or empirical estimation from the gel band intensity. Follow the manufacturer’s instructions to run 1 μ L of library on a High Sensitivity chip (or 1 μ L of each pooled library if multiple pooled libraries have been prepared in parallel). Check that the library is the correct size (the BioAnalyzer peak commonly appears slightly larger than the theoretical library size, e.g., a 271 bp library is often measured as approx. 280–300 bp) with a very small or absent “primer-dimer-adapter” off-target product peak. Off-target peaks will diminish the efficiency of the use of the “sequencing space.” If more off-target product is evident than is acceptable, conditions should be further optimized (e.g., altered cycle numbers, adjusted annealing temperatures, gel loading quantity, or re-design of problem primers). Figure 3b depicts typical successful pooled library product as assessed by BioAnalyzer.
8. Use the calculations provided by the BioAnalyzer to convert the Qubit pg/ μ L concentration reading into molarity based on the bp size of the library. The “insert” length (and, therefore, the total library element length) for a given Hi-Plex primer design affects the conversion co-efficient. For example, for a 271 bp library (approx. 100 bp of query sequence + GSPs + barcoding adapter sequence) measured to be at 496 pg/ μ L, the calculation is as follows:

BioAnalyzer calculates a pg/ μ L to pM conversion coefficient of 5.15 so the molarity equates to $496 \times 5.15 = 2554$ pM. This conversion coefficient can be used to convert between units for other libraries of this size (bp).
9. It is critical that libraries are always re-quantified by Qubit after any concentration adjustments (e.g., dilution, drying, or pooling) or if more than 24 h have elapsed between quantification and initiating sequencing steps. If insufficient quantity or concentration is yielded, troubleshooting options are drying down, extracting more library and reviewing the PCR conditions, including considerations such as cycle numbers and input DNA quantity and quality.
10. Dilute library(ies) to 2.3 nM ready for sequencing steps (*see Note 17*).

3.4 Sequencing— Library Denaturation

1. Freshly dilute 10 M NaOH to 0.1 N (10 μ L 10 M NaOH + 990 μ L molecular grade water).
2. Denature the library (concurrently, denature PhiX according to the User guide *Preparing libraries for sequencing on the MiSeq Part # 15039740*). For each library being processed, mix 6 μ L of 2.3 nM library with 6 μ L 0.1 N NaOH to achieve a denaturation volume of 12 μ L. Immediately, start a lab timer,

briefly vortex the library, and centrifuge at $280 \times g$ for 1 min. Incubate at room temperature until the timer registers 6 min. Add 588 μL of Chilled HT1 buffer to achieve a volume of 600 μL containing the library at 23 pM and NaOH at 1 mM. We typically find best performance with higher than “usual” library concentrations (hence 23 pM). For different libraries, optimal loading can be determined empirically.

3. Add 594 μL of the library to 6 μL of denatured 12.5 pM PhiX, to achieve a molar ratio of ~1% PhiX: ~99% library.
4. Store the denatured library *on ice* until it is ready to be loaded into the MiSeq reagent cartridge (*see* Subheading 3.5, step 6).

3.5 Sequencing— Reagent Cartridge Preparation

1. Thaw the MiSeq reagent cartridge by placing it in an insulated container and adding water to the marked line on the side of the reagent kit. Thaw for approx. 1 h (alternatively, the cartridge can be thawed overnight at 4 °C).
2. When the reagent cartridge has nearly thawed, thaw aliquots of Read 1 primer, Read 2 primer, and i7 read primer custom sequencing primers (each stock at 100 μM). Pipette 3.4 μL of each primer into freshly labeled tubes. These must be added to the sequencing cartridge in subsequent steps (*see* Subheading 3.5, step 7) for the sequencing chemistry to work with the present hybrid adapter design.
3. Thaw a tube of HT1 buffer (from the MiSeq reagent cartridge box) on ice.
4. Enter details on the sample sheet to reflect the sample names and identification codes and their corresponding i7 and i5 indices. Importantly, i7 barcodes are entered in the sample sheet as the reverse complement of the 8 bases in the i7 barcoding adapter primers, whereas i5 barcodes are entered as they feature in the i5 barcoding adapter primers. Care must be taken to ensure accuracy during sample sheet population to avoid errors in the downstream analysis. Ensure that all the samples have been assigned a unique ID (including any blank “samples” if copying/generating from a previous template file) to avoid an “ID incompatibility error” when the MiSeq ICS checks the sample sheet.
5. Load the MiSeq cartridge. Blot off any excess water from the reagent cartridge after removing from the water and invert the cartridge 10 times to mix the contents. Check that the reagent reservoirs are fully thawed and that there are no bubbles in the bottoms of the reservoirs. Tap the cartridge on the bench and/or flick an individual reservoir to dislodge any bubbles.
6. Use a clean 1 mL pipette tip to pierce the foil seal over the reservoir labeled “Load Samples” and dispense 600 μL of library/PhiX mixture into the reservoir.

7. Spike in custom sequencing primers to the appropriate reagent cartridge reservoirs. Use a fresh pipette tip to pierce the foil seal over reagent reservoir number 12. Use a narrow transfer pipette or glass Pasteur pipette to collect some solution from reservoir 12 into the Eppendorf containing the 3.4 μL “Read 1 primer” aliquoted earlier and mix by gentle pipetting up and down to minimize bubble formation. Gently reload the mixed contents to the bottom of reservoir 12. Sequentially, repeat this process so that “i7 read primer” is mixed into reservoir 13 and “Read 2 primer” is mixed into reservoir 14.
8. Follow the on-screen prompts to load the flow cell, PR2 and the reagent cartridge. Proceed to sequencing according to the manufacturer’s instructions.
9. Analyze the sequencing files to detect genetic variants (*see Note 18*).

4 Notes

1. Libraries generated using this protocol should also be compatible with other sequencing-by-synthesis Illumina platforms using the equivalent chemistry with minor modifications, e.g., loading quantities, reagent cartridge design, etc. NextSeq or HiSeq3000/4000 instruments require an additional custom sequencing primer due to altered sequencing steps.
2. For a high-throughput screening approach in which multiple products are pooled prior to gel extraction, this will influence the uniformity of sequencing coverage across the specimens.
3. We typically conduct 2×150 bases paired-end sequencing to enable complete overlap of reads of read-pairs—this allows highly stringent variant calling with approx. 100 bp “target” regions (target is defined here as the library insert minus GSP regions—GSPs are included in sequencing reads).
4. This relatively small size is of benefit when working with fragmented DNA and enables high-accuracy read-overlap analysis via ROVER or UNDR-ROVER [4, 5]. Working with longer inserts should also be compatible with Hi-Plex, trading amplicon-plexity and cost benefits for variant-calling accuracy (using the current example read lengths). This approach had not been trialed at the time of writing.
5. This can impact profoundly uniformity and on-target accuracy. Strong primer-dimers are particularly problematic and should be avoided where possible. Since our earlier demonstrations of Hi-Plex, we have developed more sophisticated primer design software (unpublished) and are open to collaborating with users to access this system. Depending on

the design setting, it may be necessary to separate groups of primers into separate sub-pools for use in separate Hi-Plex reactions. We have developed in-house pooling software to guide this process.

6. Using a design target primer melting temperature of 62 °C and a minimum reaction annealing temperature of 58 °C tends to work well. Primers that have much lower than targeted melting temperatures will contribute to under-representation of their corresponding amplicon. We aim to achieve an optimal balance between reaction permissivity across as many primers as possible while maintaining sufficiently stringent conditions to reduce off-target priming events. Some off-target priming events can drive major off-target effects. We have developed software (unpublished) to measure on- and off-target events associated with each gene-specific primer in the mix. Often, removal of a relatively small number of primer species (involved in major off-target effects) can dramatically improve the cleanliness of a library, on-target rate, and uniformity.
7. The current iteration of Hi-Plex is based on a hybrid TruSeq-IonTorrent adapter design for dual sequencing chemistry compatibility [9]. However, at the time of writing, IonTorrent chemistry does not lend itself to efficient barcoding using Hi-Plex. The present example illustrates Hi-Plex intended for Illumina dual index sequencing-by-synthesis chemistry using spiked-in sequencing primers to standard reagent cartridges. In theory, a wide variety of adapter designs are possible, but are as yet untested. In future work, in the interest of convenience, we will explore an all-Illumina adapter design to preclude the requirement for spiked-in sequencing primers. In the case of re-design of adapters, care should be taken to avoid primer-dimer and other off-target effects.
8. We typically use Hi-Plex in the 96-well plate format. Other plate arrangements, e.g., 384, should also work. The number and length of amplicons and the sequencing platform influence the number of specimens that should be barcoded and analyzed concurrently in a sequencing lane. The dual-indexing barcoded adapter primers listed in this chapter allow $24 \times 36 = 864$ -level specimen multiplexing, although, in theory, much higher levels should be possible. With the Illumina sequencing-by-synthesis chemistry, it is important to maintain library complexity across each base of the barcodes.
9. We have recently observed that, frequently, substantially improved results are achieved by setting up Hi-Plex PCR on ice compared with room temperature. For GSPs and universal primers, aggregate primer concentrations are listed, not the concentrations of individual primers. Phusion enzymes are sensitive to vortexing.

10. Reaction volumes could probably be scaled down to reduce reagent costs and template DNA requirements but this has not been tested and might increase the risk of evaporative loss effects.
11. This amount is based on human genomic DNA. The mass of DNA required will depend on the application. Different amounts of template DNA can be used with Hi-Plex, although the number of amplification cycles should be adjusted accordingly. Optionally, the DNA plate can be prepared in advance in a dried-down form.
12. Different target widths tend to require different cycle numbers. Different primer designs will perform optimally with different annealing/extension temperatures—we aim to provide a range of extended holds to provide the best opportunity for lower and higher GC content amplicons to complete extension with relative uniformity. In many contexts, these may well not need to be so long or varied—this would need to be determined empirically.
13. Depending on the reaction, 2 or 4 adapter cycles can give the best results, balancing yield and cleanliness. The number of amplification cycles prior can be influential.
14. Depending on the yield, it may be necessary to run more than one library lane. Optionally, a prior QC gel can be run using smaller sample loading into non-taped wells. Overloading will tend to result in excessive smearing, difficulty in resolving library bands and poor quality libraries. In our experience, “conventional” agarose gels with sufficiently wide wells are best for this purpose (versus such as E-Gels, which appear to succumb to overloading more readily).
15. Care should be taken when working with UV, ethidium bromide, and scalpel blades. The appropriate protective equipment should be worn and items disposed of in the appropriate safety containers. In our experience, ethidium bromide performs better than alternatives with respect to gel migration, image clarity, and photobleaching. Restrict UV exposure of libraries to the minimum time required.
16. Calibrated pipettes and attention to detail are key to accurate DNA quantification. Accurate quantification will have an important bearing on the relative yields of individual specimens within a library pool. Insufficient template will result in under-representation of a given specimen and over-representation of a template will result in over-sequencing and reduced overall screening efficiency. When pipetting, just touch the meniscus of a given solution to avoid carry-over. Prepare enough dye reagent for $n + 3.5$ samples (where n is the number of libraries to quantitate) to allow for S1, S2, and S2-QC

plus a slight excess to account for small pipetting errors. Qubit has an optimal range for reading accuracy, so library readings should be taken using appropriate dilutions.

17. The current protocol is written for use with four barcoding adapter PCR cycles. When two barcoding adapter PCR cycles are used, we appear to achieve cleaner library profiles at the expense higher concentrations (than products of 4 adapter cycles—which already require considerably higher concentrations to be loaded than recommended conventionally) being required to achieve high cluster densities. We suspect that the proportion of products with complete adapter sequences on both ends of a library element is smaller when two adapter cycles are employed. In both the scenarios, we suspect that a proportion of unproductive “single-ended” products co-migrate with the “double-ended” products. This is an area of current investigation—we expect that protocols will be able to be standardized based on the number of adapter cycles employed. Quantitative PCR will likely be useful to achieve a more direct indication of the library concentration to be used with the sequencer. Prior to running any “large-scale” sequencing runs, we recommend checking for performance (including cluster density) using a Nano sequencing run, for example.
18. For Hi-Plex libraries, GSP sequences are included in the reads. Depending on the primer design, overlaps of amplicons can occur. If not accounted for, these factors will confound variant detection analysis. As mentioned previously, we have published ROVER and UNDR-ROVER that are appropriate for Hi-Plex-based analyses where the reads of read-pairs have been designed to overlap completely.

Acknowledgments

The development of Hi-Plex has been supported by the Australian National Health and Medical Research Council (NHMRC) project grants 1025879 and 1108179, Cancer Council Victoria project grant 1066612 and the Victorian Life Sciences Computation Initiative (VLSCI) resource allocation VR0182.

References

1. Nguyen-Dumont T, Hammet F, Mahmoodi M et al (2015) Abridged adapter primers increase the target scope of hi-Plex. *BioTechniques* 58:33–36
2. Nguyen-Dumont T, Pope BJ, Hammet F et al (2013) A high-plex PCR approach for massively parallel sequencing. *BioTechniques* 55:69–74
3. Nguyen-Dumont T, Mahmoodi M, Hammet F et al (2015) Hi-Plex targeted sequencing is effective using DNA derived from archival dried blood spots. *Anal Biochem* 470:48–51
4. Pope BJ, Nguyen-Dumont T, Hammet F et al (2014) ROVER variant caller: read-pair overlap considerate variant-calling software applied to

- PCR-based massively parallel sequencing datasets. *Source Code Biol Med* 9:3. <https://doi.org/10.1186/1751-0473-9-3>
5. Park DJ, Li R, Lau E et al (2016) UNDR ROVER—a fast and accurate variant caller for targeted DNA sequencing. *BMC Bioinformatics* 17:165. <https://doi.org/10.1186/s12859-016-1014-9>
 6. Nguyen-Dumont T, Teo ZL, Pope BJ et al (2013) Hi-Plex for high-throughput mutation screening: application to the breast cancer susceptibility gene *PALB2*. *BMC Med Genomics* 6: 48. <https://doi.org/10.1186/1755-8794-6-48>
 7. Nguyen-Dumont T, Hammet F, Mahmoodi M et al (2015) Mutation screening of *PALB2* in clinically ascertained families from the breast cancer family registry. *Breast Cancer Res Treat* 149:547–554
 8. Hasmademail HN, Laiemail KN, Wenemail WX et al (2016) Evaluation of germline BRCA1 and BRCA2 mutations in a multi-ethnic Asian cohort of ovarian cancer patients. *Gynecol Oncol* 141(2):318–322. <https://doi.org/10.1016/j.ygyno.2015.11.001>
 9. Nguyen-Dumont T, Pope BJ, Hammet F et al (2013) Cross-platform compatibility of Hi-Plex, a streamlined approach for targeted massively parallel sequencing. *Anal Biochem* 442:127–129. <https://doi.org/10.1016/j.ab.2013.07.046>

ClickSeq: Replacing Fragmentation and Enzymatic Ligation with Click-Chemistry to Prevent Sequence Chimeras

Elizabeth Jaworski and Andrew Routh

Abstract

We recently reported a fragmentation-free method for the synthesis of Next-Generation Sequencing libraries called “ClickSeq” that uses biorthogonal click-chemistry in place of enzymes for the ligation of sequencing adaptors. We found that this approach dramatically reduces artifactual chimera formation, allowing the study of rare recombination events that include viral replication intermediates and defective-interfering viral RNAs. ClickSeq illustrates how robust, bio-orthogonal chemistry can be harnessed in vitro to capture and dissect complex biological processes. Here, we describe an updated protocol for the synthesis of “ClickSeq” libraries.

Key words ClickSeq, RNAseq, Click-chemistry, Next-generation sequencing, Flock house virus

1 Introduction

In nature, DNA is composed of long polymers of deoxyribose sugars each carrying a nucleobase that are linked together by phosphate groups. However, a number of recent studies have generated DNA and RNA with unnatural triazole-linked backbones that can approximate the natural properties of DNA or RNA [1–11]. Such molecules are generated by “click-ligating” azido- and alkyne-functionalized nucleic acid strands together via Copper-catalyzed Azide-Alkyne Cycloaddition (CuAAC), the prototypical Click-Chemistry reaction [12] (Fig. 1a) or Strain-promoted Azide-Alkyne Cycloaddition (SPAAC) [13] (Fig. 1b). The types of triazole-linkages formed are varied (Fig. 2) with some closely mimicking the native structure and base-to-base distance of natural DNA (compare structures **A** and **B** in Fig. 2), and others inserting large bulky chemical groups (e.g., structure **E** in Fig. 2). It is thus remarkable that such nucleic acid templates have been shown to be biocompatible in a number of settings: in vitro using reverse transcriptases [2] and DNA polymerases [5, 14, 15], and in vivo in *E. coli* [4, 7] and in eukaryotic cells [1].

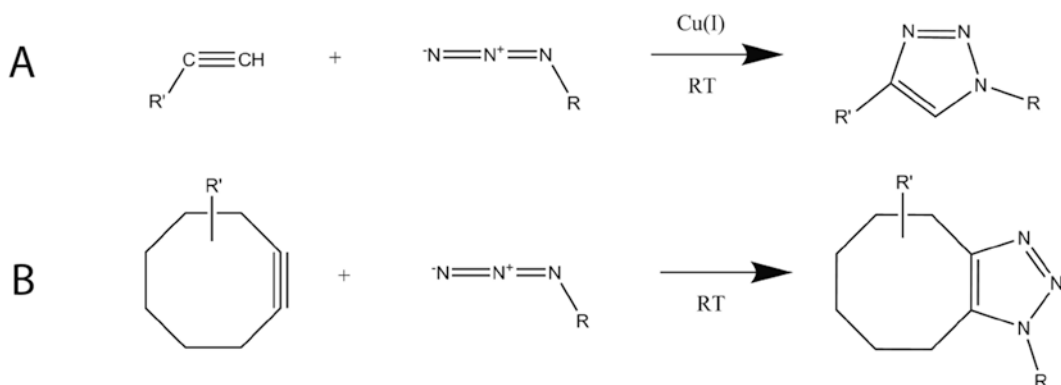


Fig. 1 Prototypical click-chemistry reactions: (a) Copper Catalyzed Alkyne-Azide Cycloaddition (CuAAC); and (b) copper-free Strain Promoted Alkyne-Azide Cycloaddition (SPAAC)

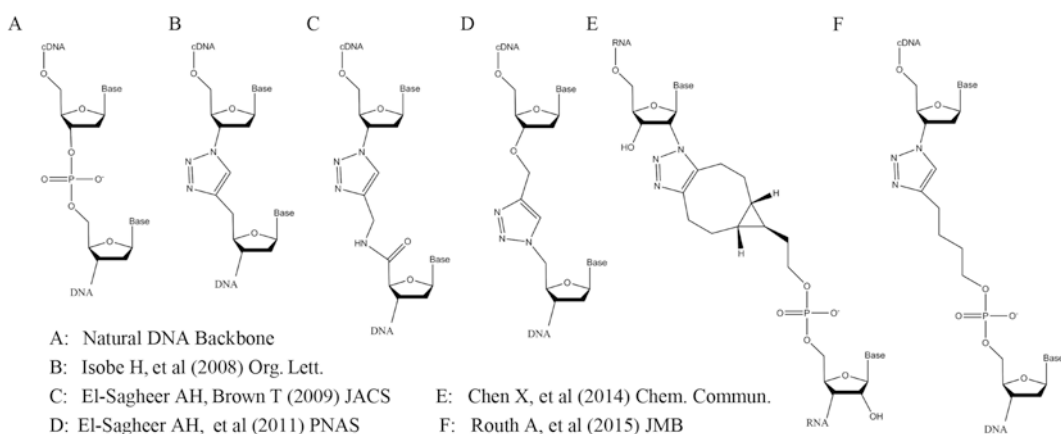
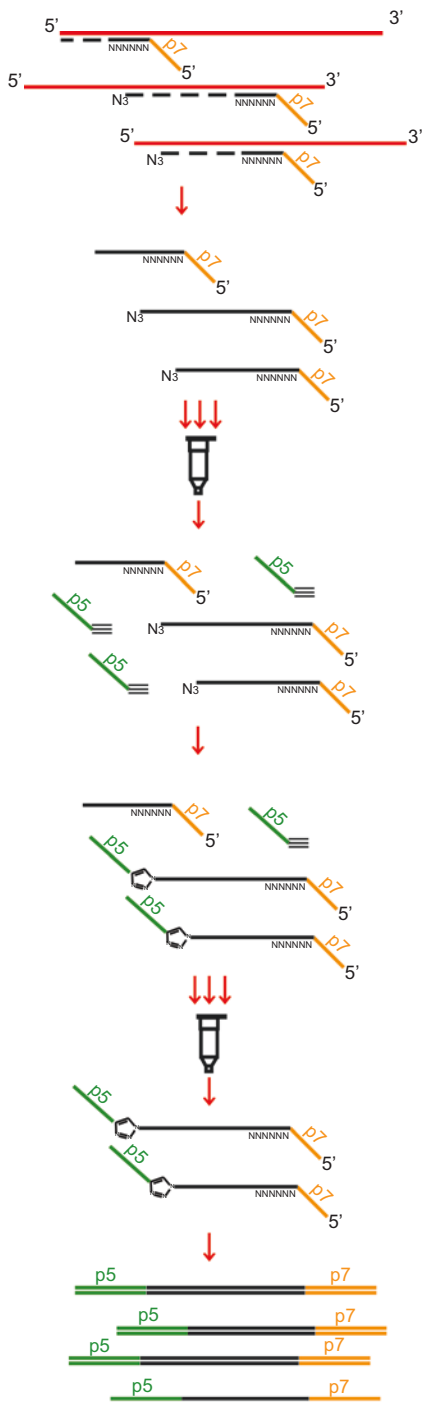
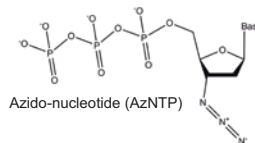


Fig. 2 A varied range of unnatural triazole-linked nucleic acids generated by click-ligation and demonstrated to be bio-compatible have been reported

Using click-chemistry to click-ligate nucleic acids together has allowed the generation of DNA templates that might otherwise be unobtainable using biological ligation, for instance, in the solid-phase synthesis of very long oligonucleotides. We recently demonstrated that click-chemistry can be used for the synthesis of Next-Generation Sequencing (NGS) libraries in a process we dubbed “ClickSeq” [15] (*see* schematic in Fig. 3). The novel innovation was to supplement randomly primed RT-PCR reactions with small amounts of 3'-azido-nucleotides to randomly terminate cDNA synthesis and release a random distribution of 3'-azido blocked cDNA fragments. These are then “click-ligated” to 5'

**Steps 3.1.1-3.1.5:**

Reverse transcription supplemented with AzNTPs and initiated from semi-random (6N) primer containing a partial p7 adaptor sequence generates a random distribution of azido-terminated cDNAs.

**Step 3.1.6:**

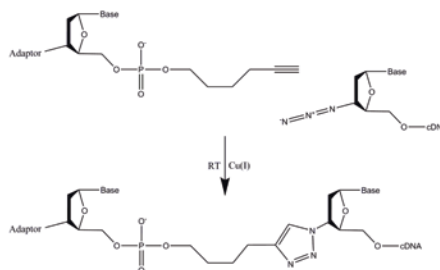
RNaseH Treatment to remove template RNA

Steps 3.2.1-3.2.4:

DNA Zymo Clean to remove RT-PCR reaction components including free azido-nucleotides.

Steps 3.3.1 - 3.3.5:

Addition of Click-Adaptor (p5) and Cu-TBTA/vitamin C catalyst mixture initiates click-ligation reaction.

**Steps 3.4.1-3.4.5:**

DNA Zymo Clean to remove copper ions, DMSO and excess adaptors to yield cleaned unnatural triazole-linked single-stranded DNA.

Steps 3.5.1 - 3.5.3:

Final PCR Amplification adds the rest of the p7 adaptor including the desired index sequences (e.g. TruSeq) and generates sufficient material for cluster generation.

Fig. 3 Schematic of the 'ClickSeq' protocol illustrating the individual steps and the Click-ligation reaction

alkyne-modified DNA adaptors via CuAAC. This generates ssDNA molecules with unnatural yet bio-compatible triazole-linked DNA backbones that can be used as PCR templates to generate RNAseq libraries.

We developed “ClickSeq” in order to address a critical limitation in NGS datasets: abundant artifactual chimeras [16]. Artifactual recombination occurs primarily due to template switching of the reverse transcriptase during RT-PCR and the inappropriate ligation of DNA/RNA fragments during cDNA synthesis [17]. In ClickSeq, azido-terminated cDNA fragments cannot provide substrates for forced copy-choice template-switching during RT-PCR as they lack a free 3' hydroxyl group required for DNA synthesis. Additionally, the azido-terminated cDNA can only be ligated to orthogonally provided alkyne-labeled DNA oligos and not to other cDNAs. Consequently, two of the main suspected sources of artifactual recombination in NGS are eliminated. When studying Flock House virus, which is known to undergo extensive recombination in vivo [18, 19], we demonstrated that artifactual recombination in NGS was reduced to fewer than three events per million reads allowing us to confidently detect rare recombination events [15, 20]. Consequently, ClickSeq allows us to explore biological systems where the rate of recombination may be much lower than what was previously detectable.

For on-going research in our lab, ClickSeq has become a routine method for making RNAseq libraries. In addition to advantages in avoiding chimeric read formation, ClickSeq does not require any template sample fragmentation. Moreover, we have recently adapted this protocol to sequence the junction of poly(A) tails and the 3'UTRs of eukaryotic mRNAs [21]. Overall, it is a simple and cost-effective procedure; once the initial click-specific reagents have been purchased the main expense is at the level of plastic-ware and PCR enzymes. This allows for the screening of multiple conditions so that a fully optimized library can be produced. Here, we describe an updated protocol for the synthesis of “ClickSeq” libraries.

2 Materials

2.1 Reverse Transcription Components

1. Deoxyribonucleotide set (dNTPs) (10 mM in water).
2. 3'-Azido-2',3'-dideoxynucleotides (AzNTPs) (10 mM in water) (Trilink Biotechnologies, San Diego, CA, USA). Reagents are stored frozen and mixed thoroughly prior to use.
3. During reverse transcription, the ratio of AzNTPs to dNTPs determines the distribution of cDNA fragment lengths generated. AzNTP:dNTP mixtures are made by making appropriate

dilutions of 10 mM AzNTPs in 10 mM dNTPs. For example, for a 1:20 10 mM AzNTP:dNTP solution add 1 μ L 10 mM AzNTPs to 20 μ L 10 mM dNTPs.

4. Reverse transcriptase: Our choice is Superscript II or III (ThermoFisher Scientific, Waltham, MA, USA) which is provided with standard reaction buffers.
5. RNaseOUT Recombinant Ribonuclease Inhibitor (ThermoFisher Scientific).
6. RNaseH.

2.2 Click-Chemistry Components

1. Click-adapter stock is resuspended in 10 mM Tris pH 8.0 and 0.5 mM EDTA at 100 μ M; working solutions of Click-adapter at 5 μ M in water.
2. Copper(II)-Tris(benzyltriazolylmethyl)amine complex (Cu-TBTA) 10 mM in 55% aq. DMSO (Lumiprobe GmbH, Hanover, Germany) or home-made.
3. 50 mM L-Ascorbic Acid is prepared by dissolving 0.44 grams powdered L-Ascorbic Acid in 50 mL water. Aliquots are dispensed into 200 μ L micro-Eppendorf tubes and stored at -20 °C. One aliquot is used fresh per experiment and discarded after use.
4. 100% DMSO.
5. 50 mM HEPES pH 7.2.

2.3 PCR Reaction

OneTaq DNA Polymerase 2 \times Master Mix with standard buffer (New England Biolabs, Ipswich, MA, USA).

2.4 Other Reagents and Equipment

1. E-Gel Precast Agarose electrophoresis system with 2% Agarose gels (ThermoFisher Scientific).
2. Blue light Transilluminator.
3. 100 bp DNA ladder.
4. Zymo DNA Clean and Concentrator-5 (Zymo Research, Irvine, CA, USA).
5. Zymo Gel DNA Recovery Kit (Zymo Research). (This kit is the same as the "Clean & Concentrator" with the addition of the agarose dissolving buffer.)
6. Qubit fluorimeter (ThermoFisher Scientific).
7. Standard Thermocyclers.
8. Standard Tabletop centrifuges.

2.5 Primers and Oligos

Primer name	Sequence	Stock soln.	Working soln.
3' Genomic Adapter-6N (partial p7 adaptor) (<i>see Note 1</i>)	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNN	100 μ M in water	Same as stock
Click-adaptor (p5 adaptor) ²	5' Hexynyl-NNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT	100 μ M in TE ¹	5 μ M in water
Universal primer short [UP_S] (p5 adaptor)	AATGATACGGCGACCACCGAG	100 μ M in TE	5 μ M in water
3' Indexing primer #1 (<i>remaining p7 adaptor</i>) ³	CAAGCAGAAGACGGCATACGAGATCGTGA <u>GTGACTG</u> GAGTTCAGACGTGT	100 μ M in TE	5 μ M in water

1. TE = 10 mM Tris pH 8.0, 1 mM EDTA
2. The Click-adaptor can be purchased from Integrated DNA Technologies (IDT). HPLC purification is required by the vendor and recommended by us.
3. Underlined portion of the 3' Indexing Primer corresponds to the TruSeq index #1. Other TruSeq indexes or other customized indexes may be used here.

3 Methods

3.1 Reverse Transcription

1. Input RNA: in principle, any input RNA can be used to generate RNAseq libraries. We have successfully sequenced viral genomic RNA, total cellular RNA, poly(A)-selected RNA, and ribo-depleted RNA. RNA should be provided in pure water, following standard precautions to avoid RNase activity. In our lab, we usually aim to provide 100 ng of RNA (*see Note 2*). No sample fragmentation is required.
2. The reverse transcription is performed using standard protocols, with the exception that the reaction is supplemented with small amounts of azido-nucleotides (AzNTPs). Set up RT-PCR reaction as follows for a 13 μ L reaction: 10 μ L of H₂O, 1 μ L of dNTP:AzNTP mixture at 10 mM (*see Notes 3 and 4*), 1 μ L of 3' Genomic Adapter-6N primer at 100 μ M, 1 μ L of RNA at 100 ng/ μ L.
3. Incubate the mixture at 95 °C for 2 min to melt RNA and immediately cool on ice for >1 min to anneal semi-random primer. This high melting temperature is tolerated as small

amounts of RNA fragmentation do not diminish efficiency of library generation.

4. Add the following at room temperature for a final reaction volume of 20 μL (*see Note 5*): 4 μL of 5 \times Superscript First Strand Buffer, 1 μL of 0.1 M DTT, 1 μL of RNase OUT, 1 μL of Superscript III Reverse Transcriptase.
5. Incubate with the following steps: 25 $^{\circ}\text{C}$ for 10 min (this step should be skipped if using a template-specific primer, *see Note 1*), 50 $^{\circ}\text{C}$ for 40 min, 75 $^{\circ}\text{C}$ for 15 min, and hold at 4 $^{\circ}\text{C}$.
6. To remove template RNA, add 0.5 μL RNase H (NEB) and incubate at 37 $^{\circ}\text{C}$ for 20 min, 80 $^{\circ}\text{C}$ for 10 min, and then hold at 4 $^{\circ}\text{C}$.

3.2 Azido-Terminated cDNA Purification

After cDNA synthesis and RNA digestion, the azido-terminated cDNA must be purified away from the AzNTPs present in the RT-PCR reaction mix. These small molecules will be in molar excess of azido-terminated cDNA by many orders of magnitude and will compete for ligation to the alkyne-modified “click-adaptor” if not completely removed. This can be achieved in a number of ways; we prefer to use the Zymo DNA clean protocol due to its ability to elute in small volumes with minimal carry-over:

1. Take the 20.5 μL RT-PCR reaction, and add 140 μL of Zymo DNA binding buffer (7:1 binding buffer:DNA).
2. Apply to silica column, and centrifuge for 30–60 s at 14,000 rpm in a tabletop microfuge, as per the manufacturer’s protocol.
3. Wash with 200 μL of ethanol-containing wash buffer and centrifuge for 30–60 s at 14,000 rpm as per the manufacturer’s protocol. Repeat for two washes.
4. Elute by centrifugation for 60 s at 14,000 rpm into fresh non-stick Eppendorf tubes using 10 μL of 50 mM HEPES pH 7.2 or water (*see Note 6*).

3.3 Click-Ligation

Following purification of the single-stranded azido-terminated cDNA, the click-ligation reaction is performed to join the 5’ alkyne-modified click-adaptor onto the 3’ end of the azido terminated cDNA. This generates a longer single-stranded cDNA with a triazole-ring and a long hexynyl linker in place of a phosphate backbone (*see Fig. 2f*).

1. First, dilute the azido-terminated cDNA in DMSO and add a large molar excess of the click-adaptor using the following volumes: 10 μL of azido-terminated cDNA (in HEPES), 20 μL of 100% DMSO (*see Note 7*), 3 μL of Click-Adapter at 5 μM in water (note: EDTA will chelate copper required in click-reaction and so must be minimized).

2. Next, generate the catalyst and accelerant mixture (for multiple samples, prepare a stock mixture): 0.4 μL of Vitamin C at 50 mM, 2 μL of Cu-TBTA in 55% DMSO.
3. Upon addition of Vitamin C, the Cu-TBTA reagent will turn from a light blue to colorless liquid, indicating the reduction of the Cu(II) ions to Cu(I). Wait 30–60 s to ensure full reduction of the copper ions (*see* **Note 8**).
4. Add 2.4 μL of the Vitamin C and Cu-TBTA mixture to each cDNA sample to initiate the click-ligation.
5. Allow the reaction to proceed at room temperature for at least 30 min (*see* **Notes 9–11**).

3.4 Click-Ligated cDNA Purification

To remove the components of the click-ligation we use the Zymo DNA clean protocol due to its ability to elute in small volumes with minimal carry-over (*see* **Note 12**):

1. The click-ligation reaction is first diluted with 60 μL water to a total volume of 100 μL prior to the addition of the DNA binding buffer in order to dilute the DMSO.
2. Take 100 μL click-ligation reaction, and add 700 μL Zymo DNA binding buffer (7:1 binding buffer:DNA).
3. Apply to silica column, and centrifuge for 30–60 s at 14,000 rpm, as per the manufacturer's protocol.
4. Wash with 200 μL ethanol-containing wash buffer and centrifuge for 30–60 s at 14,000 rpm as per the manufacturer's protocol. Repeat for two washes.
5. Elute by centrifugation for 60 s at 14,000 rpm into fresh non-stick Eppendorf tubes using 10 μL 10 mM Tris pH 7.4 or water.

3.5 Final PCR Amplification

We have screened a number of cycling conditions and have found the following to give the best results:

1. Mix at room temperature for a 50 μL reaction: 5 μL Clean Click-ligated DNA (in 10 mM Tris pH 7.4) (*see* **Note 13**), 2.5 μL of 3' Indexing Primer (1 barcode/sample) at 5 μM , 2.5 μL of Universal Primer Short [UP-S] at 5 μM , 15 μL of H_2O , 25 μL of 2 \times One Taq Standard Buffer Master Mix.
2. Cycle on a standard thermocycler using the following steps (*see* **Note 14**):
94 $^\circ\text{C}$ 1 min;
55 $^\circ\text{C}$ 30 s,
68 $^\circ\text{C}$ 10 min
[94 $^\circ\text{C}$ 30 s,
55 $^\circ\text{C}$ 30 s,

68 °C 2 min] × 17 cycles

68 °C 5 min;

4 °C ∞

3. Purify the PCR product with another Zymo DNA clean protocol (*see Note 15*):
 - (a) Take the 50 µL PCR reaction and add 250 µL of Zymo DNA binding buffer (5:1 binding buffer:DNA).
 - (b) Apply to silica column, and centrifuge for 30–60 s at 14,000 rpm, as per the manufacturer's protocol.
 - (c) Wash with 200 µL of ethanol-containing wash buffer and centrifuge for 30–60 s at 14,000 rpm as per the manufacturer's protocol. Repeat for two washes.
 - (d) Elute by centrifugation for 60 s at 14,000 rpm into fresh non-stick Eppendorf tubes using 20 µL of 10 mM Tris pH 7.4 or water.

3.6 Gel Extraction and Size Selection

1. Add 20 µL eluted cDNA library onto a 2% agarose precast prestained e-gel. For multiple samples, run empty wells in between each sample to prevent cross-contamination of final libraries. Also run a 100 bp MW ladder.
2. Run using 1–2% agarose protocol for 10 min (E-Gel iBASE Version 1.4.0; #7).
3. After run has completed, image gel on blue transilluminator and keep image for records (e.g., Fig. 4a).

a - Pre-cut Gel b - Size Selection

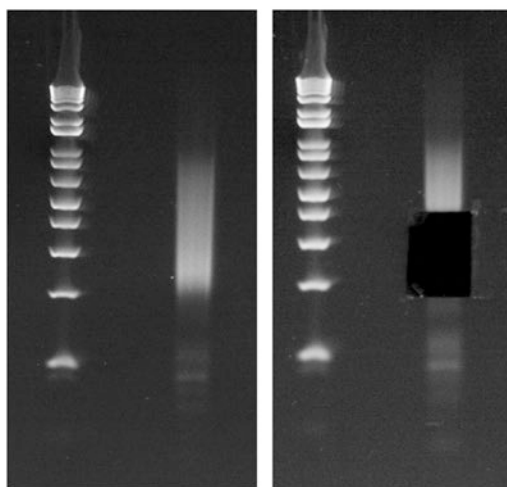


Fig. 4 The final cDNA library is analyzed by gel electrophoresis. The library should appear as a smooth smear as per the example in (a). (b) A library of the desired size is excised and an image is retained for records

4. Crack open precast gel cassette, and with a fresh/clean scalpel or razor blade, excise the desired cDNA library sizes. In ClickSeq, the total length of adapters are 126 bp. Therefore, minimum cDNA library size should be 176 bp for 1×50 bp SE Illumina. Example in Fig. 4b shows a library excised from 200 to 400 bp for a 1×75 bp SE Illumina run on a HiSeq.
5. Weigh excised gel and mix 3:1 volume for weight Zymo Agarose dissolving buffer (ADB) (e.g., 180 μ L of ADB for 60 mg agarose).
6. Incubate at 50 °C for approximately 10 min. Make sure that agarose has entirely dissolved before proceeding. Take care not to incubate at temperatures greater than 50 °C, as this may partially melt some dsDNA fragments and result in improper quantification.
7. Purify the PCR product with the Zymo DNA clean protocol: Apply melted agarose in ADB to silica column, and centrifuge for 30–60 s at 14,000 rpm, as per the manufacturer’s protocol. Wash with 200 μ L of ethanol-containing wash buffer and centrifuge for 30–60 s at 14,000 rpm as per the manufacturer’s protocol. Repeat for two washes. Elute by centrifugation for 60 s at 14,000 rpm into fresh non-stick Eppendorf tubes using 6–10 μ L of 10 mM Tris pH 7.4 or water.
8. Quantify yield of final size selected cDNA library using a QuBit fluorimeter.

3.7 Sequencing and ClickSeq-Specific Data Preprocessing

ClickSeq Libraries can be submitted for either paired-end or single-end sequencing on Illumina platforms using the adaptor sequences described here. The first read is obtained from the Illumina universal primer end (p5) end of the cDNA fragment which is the location of the triazole ring in the original cDNA. The second read starts from the indexing (p7) adaptor, which is the site of the random priming in the RT-PCR. During data preprocessing, we recommend trimming the first six nucleotides from the beginning of both the forward and reverse reads, which correspond to the random “N” nucleotides included in the sequencing adaptors (*see* Subheading 2.5). Additionally, in the forward read, we have found that there is sequence bias in the 4th to 6th nucleotides for the forward read. These nucleotides correspond to those flanking the unnatural triazole linkage. In particular, position 5 can be occupied with an “A” in up to 80% of the sequence reads. This position corresponds to the base complementary to the terminating azido--nucleotide introduced during RT-PCR, suggesting that either AzTTP is inserted more readily than the other azido-nucleotides during reverse transcription or that the click-ligation reaction favors terminal azido-thymine in the cDNA. Alternatively, it is possible that the PCR amplification step may preferentially insert an “A” opposite the triazole-linkage regardless

of the complementary base. We have not found this to adversely affect the evenness of our sequence coverage; however, future optimization may be required to eliminate any potential bias.

4 Notes

1. Template-specific primers can be used in place of semi-random primers at this step. Simply exchange the “NNNNNN” nucleotides for the sequence of choice. Similarly, Oligo-dT primers can be used, as described in [21], in order to sequence the junction of the poly(A) tails and 3' UTRs of eukaryotic mRNAs and viral RNAs. Proceed to 50 °C without initial 25 °C incubation immediately after the addition of reverse-transcription enzyme to reduce off-target amplification (*see* Subheading 3.1, step 5).
2. We have successfully generated RNAseq libraries from as little as 20 pg of starting Flock House virus RNA. However, the number of PCR cycles used for final library amplification must be greatly increased (up to 36 cycles), which will inevitably introduce sequence bias and duplication.
3. Optimal AzNTP:dNTPs ratios must be determined empirically for a given procedure; but as a general rule, 1:20 is suitable for approx. 100–200 inserts (e.g., 1×100 bp SE Illumina); and 1:35 for >250 nt inserts (e.g., 2×300 PE Illumina). We have also used nucleotide mixes that omit AzTTP when priming from poly(A) tails in order to prevent termination within the poly(A) tail [21].
4. Care must be taken when aiming to make libraries with long insert lengths and thus with large ratio of dNTPs to AzNTPs. Smaller RNA fragments will allow the reverse transcriptase to reach the end of the RNA fragment without the incorporation of an AzNTP, resulting in an un-clickable product. As a result, these fragments will be strongly under-represented in the final cDNA library.
5. At this stage, a master mix can be made. For example, if making five libraries, mix 22 µL of 5× Superscript First Strand Buffer, 5.5 µL of 0.1 M DTT, 5.5 µL of RNaseOUT, 5.5 µL of Superscript III Reverse Transcriptase, and then add 7 µL of this to each RNA/primer/dNTP mixture from Subheading 3.1, step 2. Superscript II and III Reverse Transcriptases seem to be stable during this short high-salt incubation.
6. Do not elute in the provided Zymo elution buffer which contains Tris or in a buffer that contains EDTA. Amine-rich buffers such as Tris solutions may reduce the efficiency/yield of the click-ligation reaction [22] and EDTA will collate the copper

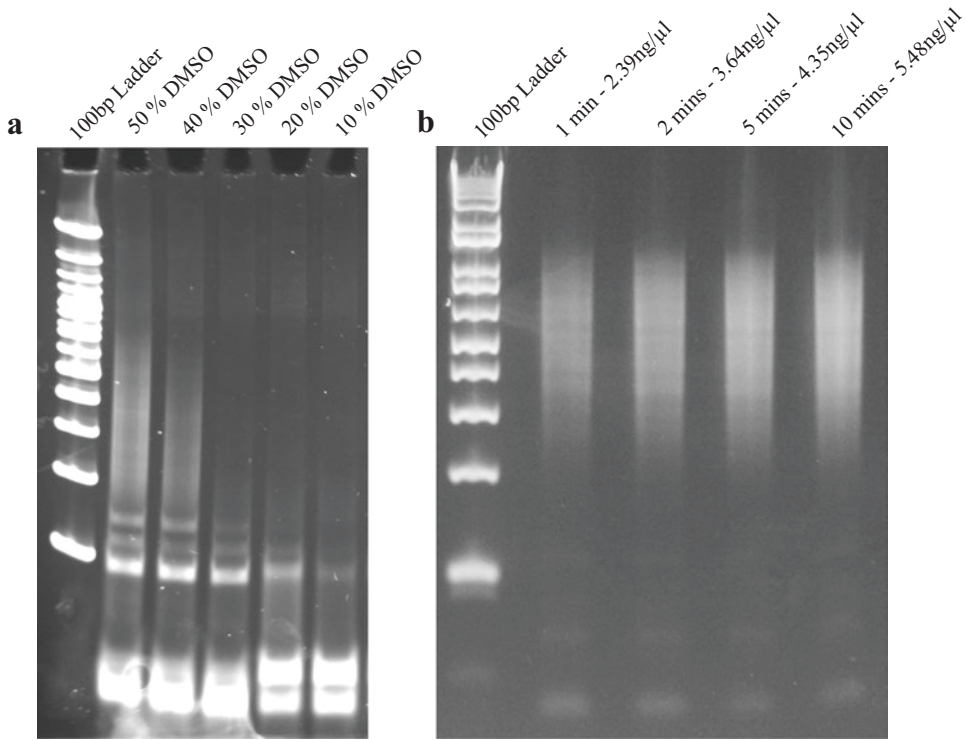


Fig. 5 Optimizations of the ClickSeq protocol are demonstrated by gel electrophoresis. **(a)** Increasing concentrations of DMSO in the Click-Ligation reaction improve the final yield of the amplified cDNA library. **(b)** Increasing the extension time

ions required for click reaction catalysis. We have found that HEPES elutes DNA from the Zymo columns well. Other buffers that are slightly alkaline (to release DNA from the silica matrix) and that are compatible with the click-ligation reaction may also be suitable (e.g., potassium phosphate).

7. To determine the optimal concentration of DMSO during the click-reaction, we performed side-by-side comparisons of libraries made using the same input azido-terminated cDNA click-ligated in the presence of 10%, 20%, 30%, 40%, and 50% DMSO. The greatest output of final library was achieved in the presence of 50% DMSO (Fig. 5a).
8. With such small volumes, the ascorbic acid reducing agent in the click-ligation reaction (whose role it is to maintain the copper catalyst in its required +1 oxidation state) is highly vulnerable to oxidation by atmospheric oxygen. Therefore, avoid introducing bubbles during pipetting and keep Eppendorf tubes closed as often as is possible. Smaller (e.g., PCR) tubes are similarly preferable.
9. The click-ligation reaction can be performed successfully at temperatures up to 90 °C. However, caution must be taken using high-temperature or extended incubation due to the

possibility of copper-mediated oxidative damage to the cDNA. This may result in an increased error-rate in base calling [23]. Moreover, we have found that extended or heightened incubation temperatures do not improve yield.

10. The click-ligation reaction can be re-supplemented with fresh catalyst/accelerant solution to ensure maximal click-ligation at regular intervals. We routinely make two total additions at 0 min and 30 min; proceeding to the cleaning step after 60 min. We have not found substantial improvement in yield with three or more additions.
11. Performing the click-reaction on the Zymo silica column itself resulted in a reduce yield, but nonetheless was feasible and results in reduction of work-flow. To “Click-on-column”: instead of eluting the azido-terminated cDNA at step Subheading 3.2, step 4, make a mixture containing the click-ligation components as detailed in step Subheading 3.3, step 1 and Subheading 3.3, step 2, except replace the azido-terminated cDNA with HEPES pH 7.2 buffer. Add 10 μ L of this mixture to the column without spinning and leave at room temperature for up to 1 h. Then, add 280 μ L of the Zymo DNA binding buffer and incubate again at room temperature for 15 min to ensure that the cDNA remains bound to the silica matrix. Next, wash the column two times in wash buffer as per the standard procedure and elute the click-ligated DNA in 10 mM Tris pH 7.4. Proceed directly to the final PCR amplification in Subheading 3.5.
12. In our original publication [15], we did not purify our click-ligated cDNA products from the components of the click-reaction. While this made for a simpler flow-through, it results in having to perform the final PCR reaction in a very large volume (200 μ L) in order to dilute away the large amounts of DMSO to acceptable levels. Additionally, without cleaning, the catalytic copper ions from the click-ligation would remain the PCR reaction mixture and may induce DNA damage due to the high cycling temperatures used during PCR [23]. Therefore, we prefer to purify the click-ligated cDNA. This may be achieved in a number of ways (e.g., EtOH precipitation, AMPure beads, etc.); here we use Zymo DNA clean columns using the standard protocols.
13. We often find it useful to amplify only half of the total purified click-ligated DNA so that a second library can later be made with fewer or more PCR cycles in case the yield of the final library is found to be inadequate or over-amplified.
14. Recently, researchers have reported that up to an 80% read-through of triazole-linked DNA templates can be achieved using non-thermostable Klenow polymerases with very long incubations [24]. To determine whether longer PCR cycling conditions would improve final library yield, we performed the PCR cycles

as described, but either with a 1 min extension time in the initial and all subsequent cycles, or 2 min, 5 min, or 10 min extension time in the initial cycle followed by 2 min extension times in all subsequent cycles (Fig. 5b). The longer extension time in the first cycle improved yield by approx. two fold.

15. PCR products can also be cleaned at this stage using AMPure beads if preferred. Add 50 μ L of PCR reaction to 50 μ L of AMPure beads and incubate at room temperature for 15 min. Wash the beads on a magnetic rack twice with 80% EtOH in water. Air-dry for 15 min. Elute library in 20 μ L of 10 mM Tris pH 7.4.

Acknowledgments

This work was supported by UTMB start-up funds and a University of Texas System Rising STARS Award to A.R.

References

1. Birts CN et al (2014) Transcription of click-linked DNA in human cells. *Angew Chem Int Ed Engl* 53(9):2362–2365
2. Chen X, El-Sagheer AH, Brown T (2014) Reverse transcription through a bulky triazole linkage in RNA: implications for RNA sequencing. *Chem Commun (Camb)* 50(57):7597–7600
3. Dallmann A et al (2011) Structure and dynamics of triazole-linked DNA: biocompatibility explained. *Chemistry* 17(52):14714–14717
4. El-Sagheer AH, Sanzone AP, Gao R, Tavassoli A, Brown T (2011) Biocompatible artificial DNA linker that is read through by DNA polymerases and is functional in *Escherichia Coli*. *Proc Natl Acad Sci U S A* 108(28):11338–11343
5. el-Sagheer AH, Brown T (2011) Efficient RNA synthesis by in vitro transcription of a triazole-modified DNA template. *Chem Commun (Camb)* 47(44):12057–12058
6. Qiu J, El-Sagheer AH, Brown T (2013) Solid phase click ligation for the synthesis of very long oligonucleotides. *Chem Commun (Camb)* 49(62):6959–6961
7. Sanzone AP, El-Sagheer AH, Brown T, Tavassoli A (2012) Assessing the biocompatibility of click-linked DNA in *Escherichia Coli*. *Nucleic Acids Res* 40(20):10567–10575
8. Isobe H, Fujino T, Yamazaki N, Guillot-Nieckowski M, Nakamura E (2008) Triazole-linked analogue of deoxyribonucleic acid ((TL) DNA): design, synthesis, and double-strand formation with natural DNA. *Org Lett* 10(17):3729–3732
9. Isobe H, Fujino T (2014) Triazole-linked analogues of DNA and RNA ((TL)DNA and (TL) RNA): synthesis and functions. *Chem Rec* 14(1):41–51
10. Fujino T et al (2011) Triazole-linked DNA as a primer surrogate in the synthesis of first-strand cDNA. *Chem Asian J* 6(11):2956–2960
11. Shivalingam A, Tyburn AE, El-Sagheer AH, Brown T (2017) Molecular requirements of high-fidelity replication-competent DNA backbones for orthogonal chemical ligation. *J Am Chem Soc* 139(4):1575–1583
12. Kolb HC, Finn MG, Sharpless KB (2001) Click chemistry: diverse chemical function from a few good reactions. *Angew Chem Int Ed Engl* 40(11):2004–2021
13. Baskin JM et al (2007) Copper-free click chemistry for dynamic in vivo imaging. *Proc Natl Acad Sci U S A* 104(43):16793–16797
14. El-Sagheer AH, Brown T (2009) Synthesis and polymerase chain reaction amplification of DNA strands containing an unnatural triazole linkage. *J Am Chem Soc* 131(11):3958–3964
15. Routh A, Head SR, Ordoukhanian P, Johnson JE (2015) ClickSeq: fragmentation-free next-generation sequencing via click ligation of adaptors to stochastically terminated 3'-Azido cDNAs. *J Mol Biol* 427(16):2610–2616

16. Gorzer I, Guelly C, Trajanoski S, Puchhammer-Stockl E (2010) The impact of PCR-generated recombination on diversity estimation of mixed viral populations by deep sequencing. *J Virol Methods* 169(1):248–252
17. Meyerhans A, Vartanian JP, Wain-Hobson S (1990) DNA recombination during PCR. *Nucleic Acids Res* 18(7):1687–1691
18. Routh A, Ordoukhanian P, Johnson JE (2012) Nucleotide-resolution profiling of RNA recombination in the encapsidated genome of a eukaryotic RNA virus by next-generation sequencing. *J Mol Biol* 424(5):257–269
19. Routh A, Johnson JE (2014) Discovery of functional genomic motifs in viruses with ViReMa—a virus recombination mapper-for analysis of next-generation sequencing data. *Nucleic Acids Res* 42(2):e11
20. Jaworski E, Routh A (2017) Parallel ClickSeq and Nanopore sequencing elucidates the rapid evolution of defective-interfering RNAs in flockhouse virus. *PLoS Pathog* 13(5):e1006365
21. Routh A et al (2017) Poly(a)-ClickSeq: click-chemistry for next-generation 3-end sequencing without RNA enrichment or fragmentation. *Nucleic Acids Res* 45(12):e112
22. Hong V, Presolski SI, Ma C, Finn MG (2009) Analysis and optimization of copper-catalyzed azide-alkyne cycloaddition for bioconjugation. *Angew Chem Int Ed Engl* 48(52):9879–9883
23. Abel GR, Calabrese ZA, Ayco J, Hein JE, Ye T (2016) Measuring and suppressing the oxidative damage to DNA during Cu(I)-catalyzed azide-alkyne cycloaddition. *Bioconjug Chem* 27(3):698–704
24. Litovchick A et al (2015) Encoded library synthesis using chemical ligation and the discovery of sEH inhibitors from a 334-million member library. *Sci Rep* 5:10916

Genome-Wide Analysis of DNA Methylation in Single Cells Using a Post-bisulfite Adapter Tagging Approach

Heather J. Lee and Sébastien A. Smallwood

Abstract

DNA methylation is an epigenetic mark implicated in the regulation of key biological processes. Using high-throughput sequencing technologies and bisulfite-based approaches, it is possible to obtain comprehensive genome-wide maps of the mammalian DNA methylation landscape with a single-nucleotide resolution and absolute quantification. However, these methods were only applicable to bulk populations of cells. Here, we present a protocol to perform whole-genome bisulfite sequencing on single cells (scBS-Seq) using a post-bisulfite adapter tagging approach. In this method, bisulfite treatment is performed prior to library generation in order to both convert unmethylated cytosines and fragment DNA to an appropriate size. Then DNA fragments are pre-amplified with concomitant integration of the sequencing adapters, and libraries are subsequently amplified and indexed by PCR. Using scBS-Seq we can accurately measure DNA methylation at up to 50% of individual CpG sites and 70% of CpG islands.

Key words DNA methylation, High-throughput sequencing, Bisulfite sequencing, Single cell, Epigenetics

1 Introduction

Methylation of cytosine bases (DNAm) is an epigenetic mark that regulates gene expression in diverse biological contexts, including cell fate specification and reprogramming, parental imprinting, repression of repetitive elements, and X-chromosome inactivation [1–4]. Development of High-throughput/Next-Generation Sequencing technologies now enables comprehensive genome-wide assessment of epigenetic modifications of mammalian cells and tissues. Regarding DNA methylation (DNAm) of cytosines residues (CpGs), the gold-standard approach to investigate the distribution and dynamic regulation of this epigenetic mark is whole genome bisulfite sequencing (WGBS) which offers single-cytosine resolution and absolute quantification of DNAm levels [5].

WGBS has now been performed on a variety of mouse and human tissues, in particular but not only, via large consortium such as ENCODE and NIH Roadmap Epigenomics [6–9].

These studies have shown that the vast majority of DNAm across tissues and cell types appear highly stable, and that DNAm variations are observed in cis-regulatory elements. While WGBS is now achievable from a wide range of starting material quantities (even from hundreds of cells [10–15]), it is limited to the analysis of bulk populations of cells. However, some genomic regions (such as enhancers) appear to display intermediate methylation at the population level; since CpG methylation is by essence binary (methylated or unmethylated), this implies a degree of heterogeneity between DNA molecules/individual cells. Therefore, the possibility of investigating the DNAm at a single-cell level could allow a better understanding of the regulatory roles of this epigenetic mark, to better comprehend its dynamic, and how it is remodeled during cell reprogramming and development.

Here, we present a protocol for Single-Cell Bisulfite Sequencing (scBS-Seq). scBS-Seq is a powerful technique as it offers single-nucleotide resolution and absolute DNAm quantification of up to half of the DNA methylome (CpG context) from unique single cells [16]. It also allows the precise measurement of 50–80% of CpG islands per cell. By merging single-cell DNA methylomes together, it is also possible to reconstruct the complete DNA methylome from extremely limited numbers of cells (10–20) [17]. This scBS-Seq protocol is based on a Post-Bisulfite Adapter Tagging approach (PBAT) [12], and is composed of 5 main steps: (1) Bisulfite treatment resulting in conversion of unmodified cytosines to uracil and DNA fragmentation; (2) Integration of the first sequencing “adapter” and pre-amplification; (3) Capture of the adapter-tagged DNA molecules; (4) Integration of the second sequencing “adapter”; (5) library amplification and indexing by PCR; (6) library QC, pooling and sequencing. scBS-Seq libraries can be prepared in 2–3 days.

2 Materials

1. RNase Away.
2. Low-binding PCR tubes.
3. Low-binding pipette tips (Starlab).
4. Tris-HCl 1 M solution, pH 8.0 (Gibco, 15568-025).
5. Sodium dodecyl sulfate solution 10% (Sigma, 71736).
6. EDTA solution 0.5 M (Sigma, E7889).
7. Proteinase K (Sigma, P4850).
8. H₂O molecular biology grade (Gibco, 10977-035).

9. Unmethylated Lambda DNA (Promega).
10. Imprint DNA Modification Kit (Sigma).
11. M-Desulphonation Buffer (Zymo Research, D5002-5).
12. Purelink PCR micro kit (ThermoFisher).
13. PCR-Grade dNTPs: dATP, dCTP, dGTP (Roche Applied Science, 11581295001).
14. High concentration Klenow exo-, 50 U/ μ L (Enzymatics, cat. no. P7010-HC-L).
15. Exonuclease I (NEB).
16. Agencourt AMPure XP beads (Beckman Coulter).
17. Absolute ethanol.
18. M-280 streptavidin Dynabeads (ThermoFisher).
19. Binding and washing (B&W) Buffer (2 \times): 10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 M NaCl.
20. KAPA HiFi HotStart DNA polymerase (KAPA Biosystems).
21. Library Quantification Kit—Illumina/Universal (KAPA Biosystems).
22. PCR thermocycler.
23. Magnetic rack for PCR tubes (e.g., ThermoFisher 492025).
24. PCR cabinet with air flow and UV light (optional).
25. 2100 Bioanalyser and high-sensitivity DNA kit (Agilent Technologies).
26. Oligo-1: 5'biotin-ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNN*N where N are random nucleotides, and * is a Thiophosphate. HPLC purification is recommended.
27. Oligo-2: 5'CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNN*N where N are random nucleotides, and * is a Thiophosphate. HPLC purification is recommended.
28. Illumina Library PCR primer PE1.0: 5'AATGATACGGCGA CCACCGAGATCTACACTCTTTCCCTACACGACGCTC TTCCGATC*T. where * is a Thiophosphate. HPLC purification is recommended.
29. iPCRtags: 5'CAAGCAGAAGACGGCATAACGAGA TXXXXXXXXGAGATCGGTCTCGGCATTCCTGCTGAAC CGCTCTTCCGAT*C (where XXXXXXXX are unique sequences (indexes), * is a Thiophosphate. HPLC purification is recommended. For more details *see* ref. 18).

3 Methods

This protocol describes the preparation of scBS-Seq libraries. It is recommended to perform a negative control (no DNA) in parallel to detect potential contaminations or other artifacts, and a positive control with 20–50 pg of DNA. Bench, pipettes and racks used should be thoroughly cleaned with DNA decontamination solution (RNase away). It is necessary to use low-DNA-binding plastic ware. A pre-PCR environment is absolutely required to limit contamination, and the use of a PCR cabinet is strongly suggested (in case, UV-treat all reagents not containing oligos or enzymes prior to use).

3.1 Cell Lysis

1. Prepare a 6.25× lysis buffer stock as follows: for 100 μL , add 6.25 μL of 1 M Tris-HCl, pH 8.0, 37.5 μL of SDS 10%, 0.625 μL of EDTA 0.5 M, 55.6 μL of H_2O . This stock can be kept at 4 °C for 4 weeks.
2. Prepare the active lysis buffer freshly by adding per cells/PCR tubes: 6 μL of H_2O , 2 μL of 6.25× lysis buffer stock, 1 μL of proteinase K, 1 μL of Lambda DNA (60 fg per μL). It is important to avoid vortexing the samples, and instead to carefully “flick” the PCR tubes with the fingers. Lambda DNA is used to assess bisulfite conversion efficiency (*see Note 1*).
3. Incubate at 55 °C for 5 min, 37 °C for 50 min, 55 °C for 5 min. This cell lysate can be stored at -20 °C for 4 weeks (*see Note 2*).

3.2 Bisulfite Conversion

1. Use the Imprint DNA modification kit, and prepare the DNA modification solution according to the manufacturer’s protocol.
2. Add 0.5 μL of balance solution to the cell lysate from Subheading 3.1 and mix by pipetting. Incubate for 10 min at 37 °C.
3. Add 50 μL of prepared DNA modification solution, mix by pipetting, and incubate as follows: 65 °C for 90 min, 95 °C for 3 min, 65 °C for 20 min and hold at 4 °C (*see Note 3*).
4. Bind the DNA to Purelink PCR micro kit columns using 240 μL of provided Binding buffer (B2). Centrifuge, discard the supernatant, and wash the column once by centrifugation using 500 μL of Wash Buffer (W1).
5. Add 100 μL of M-Desulphonation Buffer and incubate at room temperature for 15 min.
6. Centrifuge the column and wash it twice using 200 μL of Wash Buffer (W1) following the manufacturer’s recommendations.
7. Elute in 20.5 μL of 10 mM Tris-HCl, pH 8.0.
8. Proceed to Subheading 3.3.

3.3 Oligo-1 Tagging

1. Add 4.5 μL of complementary strand synthesis mix (1 μL of dNTPs 10 mM; 2.5 μL of 10 \times Blue buffer; 1 μL of oligo-1 10 μM) to the sample, and mix by pipetting.
2. Incubate at 65 $^{\circ}\text{C}$ for 3 min and cool down to 4 $^{\circ}\text{C}$.
3. Add 1 μL of Klenow exo- enzyme and mix by pipetting (*see Note 4*).
4. Incubate for 5 min at 4 $^{\circ}\text{C}$ and then rise to 37 $^{\circ}\text{C}$ at a rate of 1 $^{\circ}\text{C}$ every 15 s, and incubate at 37 $^{\circ}\text{C}$ for 30 min. Pause at 4 $^{\circ}\text{C}$.
5. Proceed to Subheading 3.4.

3.4 Pre-amplification

1. Prepare the pre-amplification mix as follows: for one sample, add 0.1 μL dNTPs (10 mM stock), 1 μL of oligo-1 (10 μM stock), 0.25 μL of 10 \times Blue buffer, 0.5 μL of Klenow exo- and 0.65 μL of H_2O .
2. Incubate samples from Subheading 3.3 at 95 $^{\circ}\text{C}$ for 1 min and cool straight to 4 $^{\circ}\text{C}$ on ice.
3. Add 2.5 μL of pre-amplification and mix by pipetting.
4. Incubate for 5 min at 4 $^{\circ}\text{C}$ and then rise to 37 $^{\circ}\text{C}$ at a rate of 1 $^{\circ}\text{C}$ every 15 s, and incubate at 37 $^{\circ}\text{C}$ for 30 min. Pause at 4 $^{\circ}\text{C}$.
5. Repeat **steps 2–4** for another three times.
6. Proceed to Subheading 3.5.

3.5 Exonuclease I Treatment and Purification

1. Complete the reaction volume to 98 μL with H_2O (*see Note 5*).
2. Add 2 μL of Exonuclease I and mix by pipetting.
3. Incubate at 37 $^{\circ}\text{C}$ for 60 min. Pause at 4 $^{\circ}\text{C}$.
4. Purify using AMPure XP beads according to the manufacturer's instructions with a bead ratio of 0.8 \times (i.e., add 80 μL of beads).
5. Elute from beads in 50 μL of 10 mM Tris-HCl, pH 8.0 and transfer to new PCR tubes.
6. Proceed to Subheading 3.6.

3.6 Capture of the Tagged DNA Strands

1. Prepare 10 μL of M-280 Streptavidin Dynabeads per sample by washing them twice with 2 \times B&W buffer according to the supplier's protocol. Resuspend the beads in 50 μL of 2 \times B&W buffer per sample.
2. Add the beads to the sample from Subheading 3.5 and incubate for 30 min at room temperature on a rotating wheel.
3. Place the PCR tubes on a magnet and wash the beads containing the tagged DNA strands twice with 100 μL of freshly prepared 0.1 M NaOH, followed by two washes with 100 μL of 10 mM Tris-HCl, pH 8.0.

4. Remove all liquid from the beads and resuspend them by pipetting with the following enzymatic reaction mix (per sample): 2 μL dNTPs (10 mM stock), 2 μL of oligo-2 (10 μM stock), 5 μL of 10 \times Blue buffer, and 39 μL of H_2O .
5. Proceed to Subheading 3.7.

3.7 Oligo-2 Tagging

1. Incubate samples from Subheading 3.6 at 95 $^\circ\text{C}$ for 1 min and cool straight to 4 $^\circ\text{C}$ on ice.
2. Add 1 μL of Klenow exo- and mix by pipetting.
3. Incubate for 5 min at 4 $^\circ\text{C}$ and then rise to 37 $^\circ\text{C}$ at a rate of 1 $^\circ\text{C}$ every 15 s, and incubate at 37 $^\circ\text{C}$ for 30 min. Pause at 4 $^\circ\text{C}$.
4. Elute from beads in 24 μL of 10 mM Tris-HCl, pH 8.0 and transfer 23 μL to new PCR tubes.
5. Proceed to Subheading 3.8.

3.8 Library Amplification and Indexing

1. Place the tubes on a PCR magnet.
2. Remove the reaction mix and wash the beads twice with 50 μL of 10 mM Tris-HCl, pH 8.0.
3. Remove the 10 mM Tris-HCl, pH 8.0 buffer and resuspend in the following PCR reaction mix (per sample): 1 μL dNTPs (10 mM stock), 10 μL of 5 \times Fidelity buffer, 1 μL of PE1.0 oligo (10 μM stock), 1 μL of iPCRtag (10 μM stock), 1 μL of KAPA HiFi Hotstart DNA polymerase, and 36 μL of H_2O .
4. Place the reaction in a PCR cyclor and incubate at 95 $^\circ\text{C}$ for 2 min, followed by 13 cycles of 94 $^\circ\text{C}$ for 80 s, 65 $^\circ\text{C}$ for 30 s, 72 $^\circ\text{C}$ for 30 s, and a final elongation at 72 $^\circ\text{C}$ for 3 min. Pause at 4 $^\circ\text{C}$ (*see Note 6*).
5. Proceed to Subheading 3.9, importantly and obviously in a different lab space (pre-PCR is not required, etc.).

3.9 Library Purification, QC, and Quantification

1. Place the samples from Subheading 3.8 on a PCR magnet.
2. Transfer the supernatant into new PCR tubes.
3. Dilute the reaction by adding 50 μL of 10 mM Tris-HCl, pH 8.0 (*see Note 5*).
4. Purify using AMPure XP beads according to the manufacturer's instructions with a bead ratio of 0.8 \times (i.e., add 80 μL of beads).
5. Elute the library in 15 μL of 10 mM Tris-HCl, pH 8.0.

3.10 Quality Control

Prior to sequencing, perform quality control using the BioAnalyzer platform according to the manufacturer's instructions (1 μL required). scBS-Seq libraries are generally characterized by a unique wide peak (~300–600 bp) (*see Fig. 1 and Note 7*). In addi-

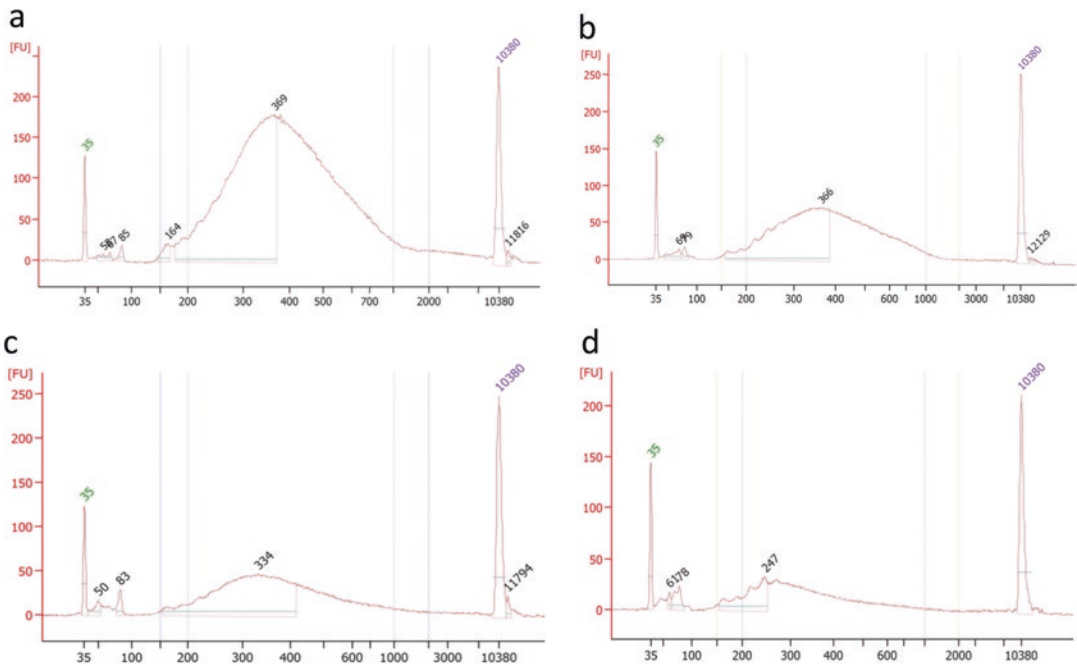


Fig. 1 Quality assessment of scBS-seq libraries using Bioanalyzer. **(a, b)** Typical Bioanalyzer profiles of good quality and successfully sequenced scBS-seq libraries generated on two separate experiments. This highlights the variability in protocol efficiency (these examples are on the extremes). **(c, d)** Below are the corresponding bioanalyzer profiles for the experimental negative controls (cell omission). Note that size distribution shift toward shorter fragments compared to single-cell samples, and the yield is much lower. It is nevertheless required to sequence negative controls

tion, use 1 μ L for quantification by qPCR using the Library Quantification Kit (KAPA biosystems) according to the manufacturer’s protocol.

3.11 Sequencing and Methylation Calls

Libraries can be sequenced on the Illumina MiSeq (for test purposes) or HiSeq. To obtain the maximum number of informative CpGs, we recommend using 200 cycle paired-end read (2×100 bp) sequencing kits. The number of reads required per sample depends obviously on the biological question asked (and the genome size); in “routine” we normally multiplex 24–48 diploid mammalian cells per lane resulting in an average of 10–25% informative CpGs. For subsequent bioinformatics analysis, we recommend that sequencing quality be first assessed using “FastQC” and poor quality sequences be trimmed using “Trim Galore!”. In addition, the first 6 bp corresponding to the oligo-1 random nucleotides can be trimmed. Subsequently, mapping and methylation calls can be performed using “Bismark” [19]; because of the PBAT strategy and pre-amplification steps, mapping should be done in single-end/non-directional mode. These three tools can be freely downloaded at the Babraham Bioinformatics website (<http://www.bioinformatics.babraham.ac.uk/projects/>).

Several reads will be present per CpGs due to the mapping strategy and the pre-amplification step. However, this does not affect the methylation calls as the very vast majority of the CpGs (>95%) are either fully methylated or fully unmethylated [16]. Samples with unusual numbers of CpGs with intermediate level of methylation compared to similar samples processed in parallel, could be the indication that multiple cells and not one cell were originally present.

4 Notes

1. Unmethylated Lambda DNA can be used to assess bisulfite conversion efficiency. Bisulfite conversion can also be assessed by measuring the DNAm levels of Cytosine in a non-CpG context. One can also investigate the mitochondrial chromosome, which should be unmethylated in most cell types. In our hands the conversion efficiency is at least 95%.
2. Cells can be collected (by FACS or mouth pipetting) directly into 9 μ L of active lysis buffer (without Proteinase K) or 6 μ L of PBS and stored at -80 °C until processing (for up to 1 month). In this case, just add the missing reagents and proceed to the incubations as described in **step 3** of Subheading **3.1**.
3. Bisulfite-converted DNA is single stranded and relatively unstable. It is best to perform the purification and Oligo-1 tagging immediately after the completion of the bisulfite treatment.
4. It is important to use a highly concentrated (HC) Klenow exo- enzyme. Using “regular” enzyme generates libraries with a much lower yield. HC Klenow exo- is also available from Sigma and NEB and both were used successfully.
5. Diluting the enzymatic reaction improves purification with AMPure XP beads by reducing the viscosity of the solution. It also improves the efficacy of ethanol washes.
6. It is important to avoid over-amplification of the scBS-seq libraries at the risk of increasing the duplication level resulting in overall loss of informative CpGs for the same sequencing depth. With this protocol, 13 cycles of PCR are normally sufficient for mouse diploid cells. For test purposes, additional PCR cycles could be performed.
7. Due to the sensitive nature of this protocol, it is frequent that traces of amplified material are detected in the negative controls. However, single-cell samples should have a higher library yield, and the library size distribution of the negative

control (as assessed on the bioanalyzer) is usually shorter. It is critical to sequence negative controls and to make sure they do not map significantly (we usually observe 3–5% mapping efficiency).

References

1. Smith ZD, Meissner A (2013) DNA methylation: roles in mammalian development. *Nat Rev Genet* 14:204–220
2. Ferguson-Smith AC (2011) Genomic imprinting: the emergence of an epigenetic paradigm. *Nat Rev Genet* 12:565–575
3. Smallwood SA, Kelsey G (2012) De novo DNA methylation: a germ cell perspective. *Trends Genet* 28:33–42
4. Seisenberger S, Peat JR, Hore TA et al (2012) Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers. *Philos Trans R Soc Lond Ser B Biol Sci* 368:20110330
5. Harris RA, Wang T, Coarfa C et al (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* 28:1097–1105
6. Bernstein BE, Stamatoyannopoulos JA, Costello JF et al (2010) The NIH Roadmap Epigenomics mapping consortium. *Nat Biotechnol* 28(10):1045–1048
7. ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* 306(5696):636–640
8. Hon GC, Rajagopal N, Shen Y et al (2013) Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat Genet* 45(10):1198–1206
9. Xie W, Schultz MD, Lister R et al (2013) Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* 153(5):1134–1148
10. Smallwood SA, Tomizawa S-I, Krueger F et al (2011) Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nat Genet* 43:811–814
11. Smallwood SA, Kelsey G (2012) Genome-wide analysis of DNA methylation in low cell numbers by reduced representation bisulfite sequencing. In: *Genomic imprinting*. Humana Press, Totowa, NJ, pp 187–197
12. Miura F, Enomoto Y, Dairiki R et al (2012) Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res* 40:e136
13. Shirane K, Toh H, Kobayashi H et al (2013) Mouse oocyte methylomes at base resolution reveal genome-wide accumulation of non-CpG methylation and role of DNA methyltransferases. *PLoS Genet* 9:e1003439
14. Kobayashi H, Sakurai T, Imai M et al (2012) Contribution of intragenic DNA methylation in mouse gametic DNA methylomes to establish oocyte-specific heritable marks. *PLoS Genet* 8:e1002440
15. Peat JR, Dean W, Clark SJ et al (2014) Genome-wide bisulfite sequencing in zygotes identifies demethylation targets and maps the contribution of TET3 oxidation. *Cell Rep* 9(6):1990–2000
16. Smallwood SA, Lee HJ, Angermueller C et al (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* 11(8):817–820
17. Nashun B, Hill PW, Smallwood SA et al (2015) Continuous histone replacement by Hira is essential for normal transcriptional regulation and de novo DNA methylation during mouse oogenesis. *Mol Cell* 60(4):611–625
18. Quail MA, Otto TD, Gu Y et al (2012) Optimal enzymes for amplifying sequencing libraries. *Nat Methods* 9:10–11
19. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for bisulfite-Seq applications. *Bioinformatics* 27:1571–1572

Sequencing of Genomes from Environmental Single Cells

Robert M. Bowers, Janey Lee, and Tanja Woyke

Abstract

Sequencing of single bacterial and archaeal cells is an important methodology that provides access to the genetic makeup of uncultivated microorganisms. We here describe the high-throughput fluorescence-activated cell sorting-based isolation of single cells from the environment, their lysis and strand displacement-mediated whole genome amplification. We further outline 16S rRNA gene sequence-based screening of single-cell amplification products, their preparation for Illumina sequencing libraries, and finally propose computational methods for read and contig level quality control of the resulting sequence data.

Key words Single amplified genomes, Fluorescence-activated cell sorting, Multiple displacement amplification, Tagmentation, Illumina genome sequencing

1 Introduction

While most of the diversity on our planet is microbial, studies have largely been limited to those microbes amenable to cultivation, greatly skewing our view of the microbial world [1]. However, the advent of cultivation-independent molecular tools, chiefly utilizing DNA sequencing, has provided a window into the phylogenetic diversity and genetic makeup of this underexplored majority. 16S rRNA gene PCR-based molecular surveys have revolutionized our view and understanding of the breadth of environmental microorganisms, and as sequencing quality and throughput have improved, so too has the genetic toolkit used to analyze microbial ecosystems. Single-cell genomics and metagenomics are now providing access to microbial genomes directly from bulk environmental samples without the constraints of cultivation. Recent studies have begun to take advantage of both approaches, producing high-quality microbial genomes while simultaneously expanding our view of the tree of life [2–4].

Single-cell genomics can be viewed as highly complementary to shotgun metagenomics, in that metagenomic datasets cast a wide net, producing a snapshot of the whole community [5], while single-cell sequencing approaches act as a scalpel used to dissect

the community and extract the genomes of specific interest [6]. However, single-cell sequencing can be challenging, as there are a number of critical steps, each of which is dependent on the successful completion of the previous step. Given the rather extensive process of going from a bulk sample to the sequence data of a single amplified genome (SAG), we outline some of the most common approaches for each of the individual steps and provide a detailed protocol.

The single-cell production workflow involves the following steps: sample preservation and preparation, single-cell isolation, cell lysis, whole genome amplification (WGA), phylogenetic screening of WGA products, library preparation, sequencing, and quality control of both reads and assembled contigs. Sample preservation and preparation are often sample specific, but critical to downstream genome quality as reported previously [7]. Single-cell isolation begins with a cell suspension step followed by isolation of individual cells using any number of different techniques including microfluidics [8–10], fluorescence-activated cell sorting [3, 11, 12], micropipetting [13], and optical tweezers [10, 14], among others. Cell lysis and whole genome amplification are the next steps in the process. The cell lysis step must lyse/disrupt as many cell types as possible while maintaining the integrity of the genomic DNA without introducing contaminants [6], and the WGA is used to produce sufficient quantities of the target genome for downstream sequencing. Many options for WGA are currently available [13, 15, 16], however isothermal multiple displacement amplification (MDA) is by far the most popular [3, 12, 17]. While significant coverage biases remain with MDA-based WGA [18], and chimeras are known to form along hyper-branched MDA DNA structures [19], other WGA methods include similar biases while increasing hands on time in the lab, ultimately increasing the potential for contamination. Following the WGA, MDA products are typically screened using 16S rRNA gene primers and Sanger sequencing to determine the identity of the isolated cells and to select a subset of interest for genome sequencing. Single-cell sequencing libraries are then prepared from the MDA product. The Illumina Nextera XT library kits have been particularly useful in the preparation of single-cell libraries as the transposase-based libraries limit hands on time by combining the fragmentation and adapter ligation steps into a single reaction. Finally, as single-cell MDA products average one chimeric junction per 20 kb of MDA sequence [19, 20], short-read technologies such as Illumina are well suited for SAG sequencing.

Single-cell genomics pipelines are inherently susceptible to contamination, as even trace amounts of contaminating DNA may be enriched during the WGA step, severely impacting the quality of the resulting SAGs. Contamination may appear from multiple sources including the samples themselves, the lab environment,

and even vendor-supplied reagents [11, 21, 22]. Furthermore, as Illumina sequencing platforms have increased in throughput, SAGs are now multiplexed within a single-sequencing run, which increases the likelihood of well to well cross contamination. Both reads and contigs (i.e., assembled datasets) should thus be assessed for contamination prior to biological data interpretation. Typically, read level decontamination is performed by screening the reads against a contaminant database composed of common contaminant sequences. There are a number of tools available to assist with read decontamination including DeconSeq [23] and various modules from the BBtools package (<https://sourceforge.net/projects/bbmap>). When SAGs are multiplexed on a single sequencing run, cross contamination can be assessed by mapping the reads of a given library/well to all assemblies in the multiplexed sequencing run, although this method requires that the multiplexed SAGs are not derived from highly similar taxa or that the taxonomically similar SAGs are removed from such cross contamination analysis. Following assembly, contig-level quality control should also be performed. Quality assurance and decontamination of assembled SAGs has traditionally been semi-manual where SAGs are screened for non-target 16S rRNA genes, abnormal k-mer frequencies, and/or variable GC content [20]; however, ProDeGe, a tool that combines composition (k-mer frequencies) and a BLAST-based screen to identify and remove questionable contigs, has enabled automated contaminant removal [24]. Finally, genome completeness can be estimated with either a universal set of single-copy marker genes or with automated software such as CheckM that calculates a set of optimal markers based on the query genome's position in a reference tree [25]. Once quality assurance has been completed, the resulting genomes are ready for phylogenetic analysis, metabolic reconstruction, and other comparative genomic analyses [26] and can be deposited into the public databases.

With the current protocol, we describe a method for obtaining genomes from individual bacteria and archaea using FACS-based cell isolation, MDA amplification, and 16S rRNA gene-based phylogenetic screening comparable to Rinke et al. [27], followed by library preparation and shotgun sequencing using the Illumina platform. We also briefly outline the bioinformatic quality assurance of SAGs from read and contig-level decontamination to genome completion estimates.

2 Materials

2.1 Sample Preparation and Preservation

1. Sample (for example, sediment or biofilm).
2. Sterile, filtered buffer solution, for example, 1× PBS (ThermoFisher Scientific, Waltham, MA, USA).
3. TE, 100×, pH 8.0.

4. Milli-Q water (EMD Millipore, Temecula, CA, USA).
5. Molecular-grade glycerol.
6. Sterile UV-treated seawater (Sigma-Aldrich, St. Louis, MO, USA).
7. Cryovials, 2 mL.
8. Microcentrifuge.
9. Vortex.
10. Centrifuge.
11. Standard light microscope.
12. Sterile cotton swabs.
13. Ultrasonic water bath.
14. Falcon tube, 50 mL.
15. Filter, 0.2 μm .

**2.2 Single-Cell
Collection Via
Fluorescence-
Activated Cell
Sorting (FACS)**

1. Ultrapure water, such as Milli-Q water, or filtered molecular biology-grade water.
2. Bleach (5% solution of sodium hypochlorite).
3. PBS liquid concentrate, 10 \times , sterile.
4. SYBR Green fluorescent nucleic acid stain (ThermoFisher Scientific).
5. Cell sorter.
6. PCR hood with UV light for decontamination.
7. Two 2-L quartz flasks for UV treatment of sheath fluid.
8. Two stir plates, stir bars for UV treatment of sheath fluid.
9. BD Falcon 40 μm nylon cell strainer.
10. Polypropylene round-bottom tubes, 5 mL.
11. Pall Acrodisc, 32-mm syringe filter with 0.1 μm Supor membrane.
12. BD Luer-Lok tip disposable syringe, 10 mL.
13. Optical microtiter plate, e.g., LightCycler Multiwell Plate 384 (Roche Diagnostics, Indianapolis, IN, USA).

**2.3 Single-Cell Lysis
and Whole Genome
Amplification by MDA**

1. Qiagen REPLI-g Single Cell Kit (Qiagen, Valencia, CA, USA).
2. SYTO 13, 5 mM (ThermoFisher Scientific).
3. Bleach (5% solution of sodium hypochlorite).
4. Spectraline XL-1500 UV Cross-linker.
5. PCR hood with UV light for decontamination.
6. Plate reader with temperature control, or a real-time thermocycler, e.g., LightCycler 480 (Roche Diagnostics, Indianapolis, IN, USA).
7. Eppendorf Safe-Lock microcentrifuge tubes, 1.5 mL.

2.4 Phylogenetic Screening

1. Ultrapure water, such as Milli-Q water, or filtered molecular biology-grade water.
2. SsoAdvanced SYBR Green Supermix (Biorad, Pleasanton, CA, USA).
3. Primer set of choice, for example 16S rRNA gene universal primer set (926wF/1392R primer, 10 μ M each).
4. ExoSAP-IT (Affymetrix, Santa Clara, CA, USA).
5. Standard thermocycler, or a real-time thermocycler.
6. Plate shaker.
7. Optical microtiter plate.

2.5 Sequencing of Single-Amplified Genomes (SAGs)

1. Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA, USA).
2. Nextera XT Index Kit (Illumina).
3. Agencourt AMPure XP (Beckman Coulter, Carlsbad, CA, USA).
4. Nuclease-free water.
5. Ethanol, 200 proof, molecular biology-grade.
6. Bioanalyzer High Sensitivity DNA Kit (Agilent Technologies, Santa Clara, CA, USA).
7. Standard thermocycler.
8. Agilent Bioanalyzer 2100 (Agilent Technologies).

3 Methods

3.1 Sample Preparation

Sample processing and preservation differ depending on the type of sample. With the current protocol, we provide a recommended workflow (Fig. 1) for the production of SAGs from two commonly surveyed sample types: a soil sample (Subheading 3.1.1) and a biofilm sample (Subheading 3.1.2).

3.1.1 Sediment or Soil Sample

1. For a sediment or soil sample, thoroughly mix 5 g of sediment or soil with 10–30 mL sterile buffer in a 50 mL falcon tube. For soils and freshwater sediments, use 1 \times PBS as buffer. For marine sediments, use sterile-filtered seawater as buffer.
2. Vortex at 14,000 rpm for 30 s to dislodge microbes from soil.
3. Centrifuge at 2000 $\times g$ for 30 s to remove large particles.
4. Collect the supernatant and proceed to Subheading 3.2.

3.1.2 Biofilm Sample

1. Using a sterile cotton swab, collect a sample of biomass into microcentrifuge tubes already containing a sterile-filtered buffer solution (e.g., 1 \times PBS or seawater).
2. Sonicate the sample in the tube in an ultrasonic water bath for 10 min.

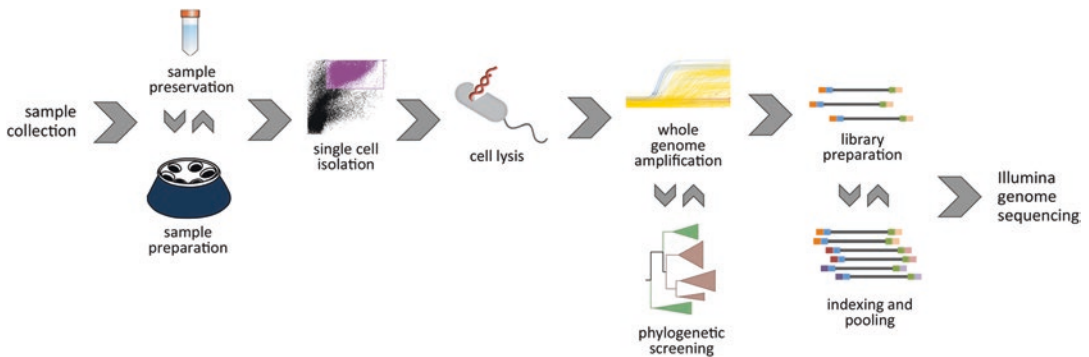


Fig. 1 Workflow for the preparation of sequencing libraries from environmental single cells. Environmental samples are typically preprocessed and cryopreserved. Single cells are then isolated using high throughput FACS, followed by lysis and MDA based whole genome amplification. MDA generated products are used as template in 16S rRNA gene based phylogenetic screening and identification. Finally, based on the results of the phylogenetic screen, a subset of single-cell MDA products is chosen for genome sequencing, following Illumina's tagmentation based Nextera XT library preparation protocol. Individual libraries may be sequenced on the MiSeq platform or pools of many barcoded libraries on higher throughput platforms such as NextSeq

3. Further shake the tube by hand for an additional 5 min.
4. Examine the sample under a microscope to ensure sufficient cell separation. If necessary, repeat **steps 2** and **3** above.

3.2 Sample Preservation

For best results and to minimize cell lysis during freeze-thawing, we recommend to cryo-protect the environmental sample and store at -80°C .

1. Cryopreservation of cells should be done with sterile-filtered glycerol. Make Gly-TE stock by mixing 20 mL of 100 \times TE (pH 8.0), 60 mL of DIW (deionized water), and 100 mL of molecular-grade glycerol (*see Note 1*). Mix solution by vortexing thoroughly and pass through a 0.2 μm filter.
2. Transfer 100 μL of prepared GlyTE stock and 1 mL of sample solution to a sterile cryovial. Mix well by inverting.
3. Prepare several replicate vials for each sample. Store in liquid nitrogen or at -80°C .

3.3 Preparation of FACS for Sterile Sort

As single-cell sorting and MDA are highly susceptible to exogenous contamination, it is essential to have a clean cell sorter with sterile sheath fluid. Briefly, fluidic lines in the cell sorter can be decontaminated by running bleach through the lines, and sheath fluid can be sterilized by UV treatment. This decontamination process is sufficient for a clean sorting process, provided it is performed before every run.

1. Within a clean hood, prepare 4 L of 1 \times PBS in two 2 L quartz flasks. Add magnetic stir bars to the flasks and place over plates

to begin stirring. Also include the empty sheath fluid tank and lid and arrange such that UV light will reach the inner surfaces. Begin the overnight UV exposure (at least 16 h). After UV exposure is complete, transfer the sterile sheath fluid to the sterile tank within the clean hood (*see Note 2*). Reserve at least 10 mL of clean sheath fluid for later use while sorting.

2. Prepare a second sheath tank with 1 L of 10% bleach (0.5% sodium hypochlorite final concentration) and run through the cell sorter for 2 h to decontaminate fluidic lines.
3. Dispose of any remaining bleach and rinse the sheath tank with sterile water. Run 1 L sterile water through the cell sorter for 30 min to rinse fluidic lines.
4. Install the dedicated clean tank with the sterile sheath fluid and begin running.
5. Prior to sorting, sterilize microtiter plates by UV treatment for 10 min. Do not seal or cover plates during UV treatment. To each well, add 2 μ L of UV-treated, 1 \times PBS (*see Note 3*). Cover with optical seal and perform an additional UV treatment for 10 min.

3.4 Cell Separation by Flow Cytometry

Due to the large diversity of types of environmental samples and variation in technical operation of cell sorters, we here outline a general protocol for sorting single cells from environmental samples.

1. To avoid clogging the nozzle of the cell sorter, filter each environmental sample through an appropriately sized filter, relative to the nozzle size (e.g., use a 20 μ m filter when working with a 70 μ m nozzle).
2. Stain the cells in the sample with 1 \times SYBR green at 4 $^{\circ}$ C for 15 min in the dark (*see Note 4*).
3. Run the stained sample through the cell sorter and target the selected population with a sort gate (*see Note 5*).
4. Sort the targeted cell population into the UV-treated optical microtiter plates containing 2 μ L 1 \times PBS per well (*see step 5* in Subheading 3.3). Any exogenous extracellular DNA that may be sorted with a single cell could present as contamination in downstream WGA. A two-step sort may be utilized to dilute this DNA (*see Note 6*). We recommend sorting into 96-well or 384-well plates and including positive and negative controls (*see Note 7*).
5. Seal plates with sterile foil and store at -80° C.

3.5 Single-Cell Lysis

The Qiagen REPLI-g Single Cell Kit is used for both single-cell lysis and WGA steps. The WGA is based on the Phi29 catalyzed MDA reaction. Refer to the manufacturer's instructions for a

detailed protocol. Note that the recommended total reaction volume of 50 μL can be reduced by up to one-fifth without any adverse effects. As an example, we here outline a 25 μL reaction.

1. Before performing any work, wipe down and clean hood surfaces, pipettes, and equipment with 10% bleach. UV the clean hood for 60 min with equipment inside (*see Note 8*).
2. Prepare lysis Buffers DLB and D2 according to the manufacturer's instructions. To each well containing a sorted single cell in 2 μL 1 \times PBS, add 1.5 μL Buffer D2. Mix thoroughly by tapping and spin down at 1000 $\times g$ for 1 min. Incubate for 10 min at 65 $^{\circ}\text{C}$.
3. Add 1.5 μL of Stop Solution to each well. Mix thoroughly by tapping and spin down at 1000 $\times g$ for 1 min. Store lysed cells (total reaction volume of 5 μL) briefly at 4 $^{\circ}\text{C}$ while preparing the WGA amplification master mix.

3.6 Whole Genome Amplification

For single-cell genome amplification, we recommend using the Qiagen REPLI-g Single Cell Amplification Kit in conjunction with SYTO staining for real-time amplification monitoring for the purpose of quality control (Fig. 2). Further, the recently developed WGA-X method is recommended for high GC templates [28].

1. Thaw the Qiagen REPLI-g Single Cell Amplification Kit reagents according to the manufacturer's instructions. Briefly, thaw the Phi29 DNA Polymerase on ice. Thaw all other reagents at room temperature.

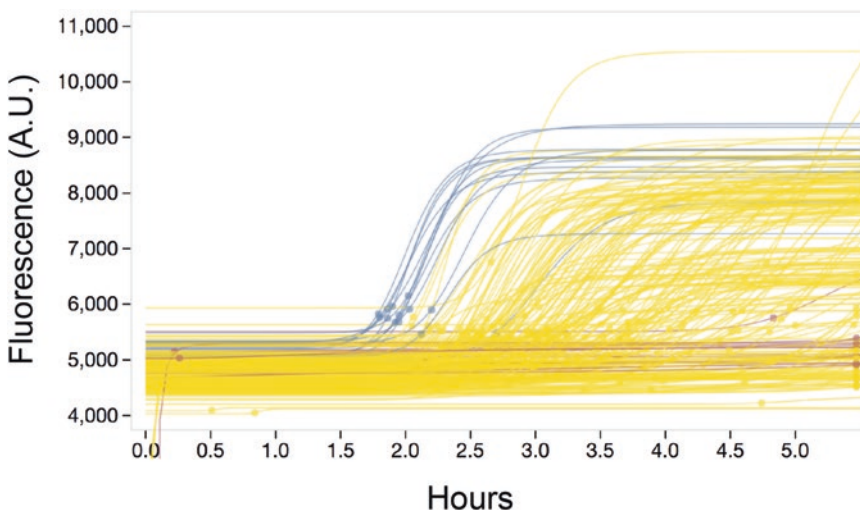


Fig. 2 Real-time kinetics for the whole genome amplification of single cells and controls in a 384-well plate. *Blue lines* show the amplification kinetics of the positive controls (100 cells sorted into a single well), *yellow lines* show the amplification kinetics of sorted single cells and *red lines* represent negative controls (no cells sorted)

2. Prepare enough master mix for all the reactions, adding Phi29 DNA Polymerase at the end. Each reaction should contain: 4.5 μL water (provided in the Qiagen kit, specifically for use in single-cell genomics), 14.5 μL Reaction Buffer, 1 μL Phi29 DNA Polymerase. Combine the water and Reaction Buffer first and mix thoroughly by vortexing before the addition of the polymerase last. Mix by vortexing and spin down (*see Note 9*).
3. To track amplification of single cells in real time, add SYTO 13 to the master mix for a final concentration of 0.5 μM (*see Note 10*). Mix by vortexing and spin down.
4. To each well containing 5 μL of lysed cells, add 20 μL of MDA master mix, for a total reaction volume of 25 μL . Cover the plate with an optical seal and centrifuge at $1000 \times g$ for 1 min (*see Note 11*).
5. Incubate the plate at 30 $^{\circ}\text{C}$ for 2–5 h in a real-time thermocycler or plate reader, e.g., the Roche LightCycler 480 (*see Note 12*).
6. Heat-inactivate the Phi29 DNA Polymerase by incubating at 65 $^{\circ}\text{C}$ for 10 min. MDA products can be stored for multiple months at -80°C (*see Note 13*).

3.7 Phylogenetic Screening

Phylogenetic screening is not mandatory, but useful for the accurate identification of potential target cells for genome sequencing. We here outline the use of universal 16S rRNA gene primers, but primers for specific functional genes may be used as well.

1. Make a 1:100 dilution of the MDA product using nuclease-free water. Mix dilution extremely thoroughly (*see Note 14*). Hand-pipetting is most effective, or alternatively, use a plate-shaker for 15 min at the maximum setting.
2. Transfer 2 μL of diluted MDA product as template to an optical microtiter plate.
3. Thaw the PCR reagents on ice: SsoAdvanced SYBR Green Supermix, 10 μM 926wF primer, 10 μM 1392R primer (*see Note 15*).
4. Prepare enough master mix for all PCR reactions. Each reaction should contain: 2.6 μL nuclease-free water, 5 μL SsoAdvance SYBR Green Supermix (2 \times), 0.2 μL 926wF primer (10 μM), 0.2 μL 1392R primer (10 μM). Mix by vortexing and spin down.
5. To each well containing 2 μL of diluted MDA product as template, add 8 μL of PCR master mix, for a total reaction volume of 10 μL . Cover the plate with an optical seal and centrifuge at $1000 \times g$ for 1 min.
6. Refer to the manufacturer's protocol when selecting a cycling program. Adding a melt curve step to the cycling program will aid in analyzing and choosing PCR products for downstream

sequencing. Begin amplification of PCR products in a real-time thermocycler.

7. Purify and sequence PCR amplicons to identify WGA products. Prior to Sanger sequencing, PCR products are purified using ExoSAP-IT according to the manufacturer's instructions. Sanger sequencing is then performed on the clean PCR products, followed by sequence analysis.

3.8 Library Preparation and Sequencing of SAGs

The Illumina Nextera XT DNA Sample Preparation Kit is used to prepare indexed paired-end libraries from single-cell MDA products for next-generation sequencing. Briefly, libraries are prepared using a single-step reaction where the transposase simultaneously fragments and ligates Illumina sequencing adapters in a single 5 min reaction. This library preparation is cost-effective and Illumina is currently the most prevalent next-generation sequencing platform.

1. Before performing any work, wipe down and clean hood surfaces, pipettes, and equipment with 10% bleach (*see Note 16*).
2. Thaw reagents for tagmentation on ice. Mix by inverting gently and spin down. To sterile microcentrifuge tubes, add: 10 μ L of Tagment DNA buffer, 5 μ L of input DNA (1 ng total), and 5 μ L of Amplicon Tagment Mix.
3. Mix by pipetting and spin down. Incubate on a thermocycler at 55 °C for 5 min and hold at 10 °C.
4. Once the thermocycler reaches 10 °C, immediately neutralize the reaction by adding 5 μ L of Neutralize Tagement Buffer. Mix by pipetting and spin down. Incubate at room temperature for 5 min.
5. Thaw the appropriate indexes and the Nextera PCR master mix. To each reaction, add 15 μ L of PCR master mix, 5 μ L of Index 1, and 5 μ L of Index 2 (*see Note 17*). Mix by pipetting and spin down. PCR amplify according to the manufacturer's protocol.
6. Perform cleanup of PCR product by using a 1.8 \times ratio of AMPure XP beads (*see Note 18*). For example, add 90 μ L of beads to each 50 μ L PCR reaction. Mix by vortexing and quick spin.
7. Shake samples at 1800 rpm for 2 min, then incubate at room temperature for 5 min.
8. Place tubes on a magnetic stand to pellet the beads (*see Note 19*). Carefully remove and discard the supernatant once the solution has become clear. Make sure not to remove any beads.
9. Without removing tubes from the magnetic stand, add 200 μ L of freshly prepared 80% ethanol to each sample. Let samples sit for 30 s. Remove and discard ethanol. Repeat, for a total of two 80% ethanol washes (*see Note 20*).

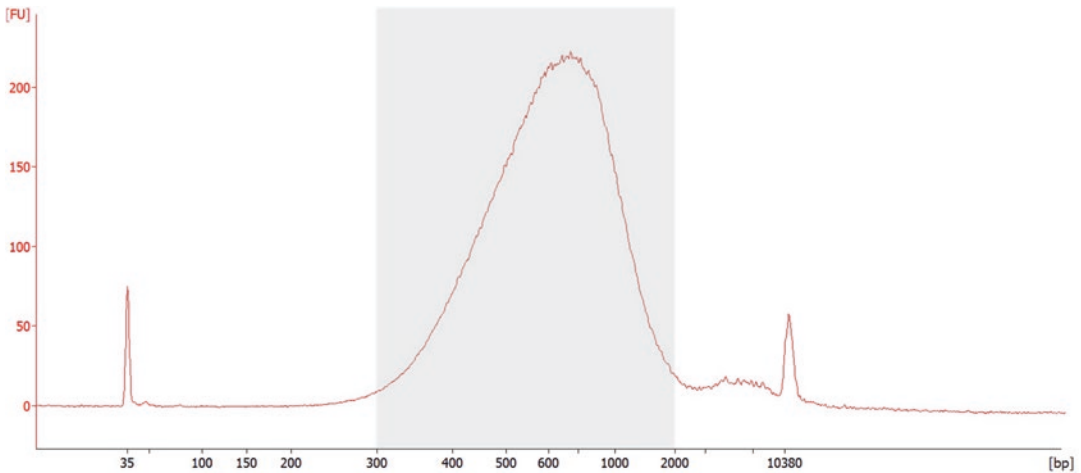


Fig. 3 Library QC of a pool of barcoded Nextera XT single-cell libraries on Agilent HS Bioanalyzer. 35 bp and 10,380 bp peaks represent markers used as internal standards. When using the protocol, library insert size distributions typically ranges from 300 to 2000 bp

10. Leave tube-caps open and let the samples air-dry on the magnetic stand for 15 min until residual 80% ethanol has evaporated.
11. Add 52 μL of Resuspension Buffer to each tube. Resuspend the beads and mix thoroughly.
12. Shake the samples at 1800 rpm for 2 min, then incubate at room temperature for 2 min.
13. Place the tubes on the magnetic stand to pellet the beads. Once the liquid is clear, carefully transfer 50 μL of eluted PCR product to a new tube.
14. Perform quality control on the final library. Run 1 μL of library on the Agilent 2100 Bioanalyzer using the High Sensitivity Kit for size and quantification (Fig. 3) (*see Note 21*).
15. Barcoded libraries can be pooled in equimolar ratios according to Illumina's pooling protocols and procedures for sequencing on the Illumina platform. For library concentrations, refer to the Illumina platform-specific protocol.
16. Sequence individual libraries or library pools using the Illumina platform, such as MiSeq or NextSeq, according to the manufacturer's protocol. NextSeq is specifically recommended for library pools [28].

3.9 Assembly of SAG Sequences and Data Quality Assurance

To ensure high-quality genomes, a few steps need to be taken before data are useable for downstream biological inference. Contaminant sequences can be identified and removed at both the read and contig (post-assembly) levels. At the read level, data are

typically removed if they match sequences from a lab specific contaminant database. Once the reads have been screened for contaminants, the decontaminated reads are used as input for a SPAdes assembly [29]. SPAdes has been specifically developed to deal with the highly uneven coverage produced during the MDA step, as SPAdes uses multiple coverage cutoffs, resulting in a larger fraction of useable data, ultimately yielding more complete assemblies. After the assembly, an additional round of contaminant screening should be performed at the contig level. Once potential contaminants have been removed, the resulting SAGs should be analyzed for genome completeness, and each of the abovementioned steps should be documented as metadata during public database submission.

1. Perform quality trimming and adapter trimming using any one of the currently available tools such as cutadapt (<http://code.google.com/hosting/moved?project=cutadapt>), Trimmomatic [30], BBDuk.sh from the BBDuk package (<http://sourceforge.net/projects/bbtools>) package package, among others.
2. Screen reads for contaminants against either premade contaminant databases such as the one employed in DeconSeq [23] or assemble an in-house contaminant database and filter out contaminant reads using BBDuk.sh from the BBDuk package (<http://sourceforge.net/projects/bbtools>).
3. Use cleaned reads as input to SPAdes [29] for denovo genome assembly.
4. Check assembled contigs for additional contaminants either manually (<https://img.jgi.doe.gov/er/doc/SingleCellDataDecontamination.pdf>) or by taking an automated approach using a tool like ProDeGe, which takes a feature-based (k-mer) approach combined with BLAST-based screening to identify suspect contigs in an assembly [24].
5. Check resulting QC'd and largely contaminant free genomes for estimated completeness and a final screen for contaminants using either a universal set of single copy markers or by implementing an automated strategy as recently outlined in CheckM [25] (*see Note 19*).

4 Notes

1. Glycerol is extremely viscous and is most accurately transferred via syringe. Store Gly-TE stock at -20°C .
2. We suggest dedicating one tank to clean sheath fluid for single-cell genomics use.
3. The amount of $1\times$ PBS added to the plate prior to sorting is determined by the total MDA reaction volume chosen (Subheading 3.5).

4. Minimize exposure of SYBR green to direct light.
5. For a wide variety of environmental samples, it can be difficult to define the target population. For example, sediment samples generally show a high background signal that may interfere with locating a distinct cell population. Other, low-biomass environments may not have a high enough cell density to easily identify a population. Some possible solutions include utilizing different nucleic acid stains or applying different sample preparation methods.
6. Begin by sorting at least 10,000 target cells into 1 mL of UV-treated 1× PBS. Backflush the sample line for 1 min, run 10% bleach solution through the sample line for 5 min, and backflush for an additional 5 min. Sort this “pre-sorted” population of cells into the UV-treated plates containing 2 μ L 1× PBS per well.
7. Positive controls are rows or columns where 10–100 cells are sorted in each well, and negative controls are rows or columns in which no cells are sorted into the corresponding wells (Fig. 2).
8. Because single-cell genomics processes are so highly susceptible to outside contamination, it is important to reserve a dedicated space for single-cell work, with separate equipment and to be maintained as sterile as possible.
9. Do not let the MDA master mix warm to a temperature above 30 °C. The Phi29 enzyme becomes inactivated at temperatures above 30 °C.
10. Before adding to the master mix, minimize exposure of SYTO13 to direct light.
11. Centrifuge the plate thoroughly to ensure that no bubbles remain in the wells. The presence of bubbles may interfere with the real-time readings measured by the instrument.
12. If directly monitoring the amplification reaction in real-time, incubation can be cut short when the negative controls begin to amplify.
13. Due to the hyper-branched nature of MDA products, MDA DNA is difficult to homogenize after a freeze-thaw cycle. If possible, proceed directly to phylogenetic screening.
14. MDA product is extremely viscous. It is important to thoroughly mix both the MDA product (pre-dilution) and the MDA dilution as well.
15. Following the manufacturer’s instructions, minimize exposure of SsoAdvance SYBR Green Supermix to direct light.
16. Prevent PCR contamination of pre-PCR processes by physically separating lab spaces where pre and post-PCR processes are performed.
17. If pooling multiple libraries together for a sequencing run, refer to the manufacturer’s protocol to select the correct

indexes for each sample, dependent on the number of libraries being pooled.

18. In preparation for bead cleanup, follow the manufacturer's instructions on best handling practices: let the beads equilibrate to room temperature before use, always vortex the beads thoroughly each time before use, and prepare fresh 80% ethanol solution before each use.
19. The pelleting process may take up to 5 min.
20. Be sure to remove all residual 80% ethanol from each sample. The tubes should remain on the magnetic stand throughout washing.
21. A typical library will have an insert size distribution of ~300–2000 bp (Fig. 3).
22. Data can be annotated and further analyzed in the IMG system (<http://img.jgi.doe.gov/>) [31].

Acknowledgment

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231. We would like to thank Bill Andreopoulos for his assistance in preparing Fig. 2.

References

1. Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 59:143–169
2. Eloe-Fadrosh EA, Ivanova NN, Woyke T, Kyrpides NC (2016) Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat Microbiol* 1:15032
3. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu W-T, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499:431–437
4. McLean JS, Lombardo M-J, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, Vyahhi N, Hall AP, Yang Y, Dupont CL, Ziegler MG, Chitsaz H, Allen AE, Yooseph S, Tesler G, Pevzner PA, Friedman RM, Nealson KH, Venter JC, Lasken RS (2013) Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proc Natl Acad Sci U S A* 110:E2390–E2399
5. Thomas T, Gilbert J, Meyer F (2012) Metagenomics – a guide from sampling to data analysis. *Microb Inform Exp* 2:3
6. Stepanauskas R (2012) Single cell genomics: an individual look at microbes. *Curr Opin Microbiol* 15:613–620
7. Clingenpeel S, Schwientek P, Hugenholtz P, Woyke T (2014) Effects of sample treatments on genome recovery via single-cell genomics. *Isme J*:1–4. <https://doi.org/10.1038/ismej.2014.92>
8. Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova NN, Martin HG, Szeto E, Platt D, Hugenholtz P, Relman DA, Quake SR (2007) Dissecting biological ‘dark matter’ with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci U S A* 104:11889–11894
9. Gawad C, Koh W, Quake SR (2016) Single-cell genome sequencing: current state of the science. *Nat Rev Genet* 17:175–188
10. Lo S-J, Yao D-J (2015) Get to understand more from single-cells: current studies of microfluidic-based techniques for single-cell analysis. *Int J Mol Sci* 16:16763–16777
11. Stepanauskas R, Sieracki ME (2007) Matching phylogeny and metabolism in the uncultured

- marine bacteria, one cell at a time. *Proc Natl Acad Sci U S A* 104:9052–9057
12. Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM, Luo H, Wright JJ, Landry ZC, Hanson NW, Thompson BP, Poulton NJ, Schwientek P, Acinas SG, Giovannoni SJ, Moran MA, Hallam SJ, Cavicchioli R, Woyke T, Stepanauskas R (2013) Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci U S A* 110: 11463–11468
 13. Zong C, Lu S, Chapman AR, Xie XS (2012) Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* 338:1622–1626
 14. Landry ZC, Giovannoni SJ, Quake SR, Blainey PC (2013) Optofluidic cell selection from complex microbial communities for single-genome analysis. *Methods Enzymol* 531:61–90
 15. Hoesjmakers W, Bártfai R, François K-J, Stunnenberg HG (2011) Linear amplification for deep sequencing. *Nat Protoc* 6:1026–1036
 16. Duhaime MB, Deng L, Poulos BT, Sullivan MB (2012) Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environ Microbiol* 14:2526–2537
 17. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, Follows MJ, Stepanauskas R, Chisholm SW (2014) Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* 344:416–420
 18. Marine R, McCarren C, Vorrasane V, Nasko D, Crowgey E, Polson SW, Wommack KE (2014) Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. *Microbiome* 2:3
 19. Lasken RS, Stockwell TB (2007) Mechanism of chimera formation during the multiple displacement amplification reaction. *BMC Biotechnol* 7:19
 20. Woyke T, Xie G, Copeland A, González JM, Han C, Kiss H, Saw JH, Senin P, Yang C, Chatterji S, Cheng J-F, Eisen JA, Sieracki ME, Stepanauskas R (2009) Assembling the marine metagenome, one cell at a time. *PLoS One* 4:e5299
 21. Blainey PC (2013) The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol Rev* 37:407–427
 22. Woyke T, Xie G, Copeland A, González JM, Han C, Kiss H, Saw JH, Senin P, Yang C, Chatterji S, Cheng J-F, Eisen JA, Sieracki ME, Stepanauskas R (2011) Decontamination of MDA reagents for single cell whole genome amplification. *PLoS One* 6:e26161
 23. Schmieder R, Edwards R (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One* 6:e17288
 24. Tennessen K, Andersen E, Clingenpeel S, Rinke C, Lundberg DS, Han J, Dangl JL, Ivanova NN, Woyke T, Kyrpides N, Pati A (2015) ProDeGe: a computational protocol for fully automated decontamination of genomes. *ISME J*. <https://doi.org/10.1038/ismej.2015.100>
 25. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. <https://doi.org/10.7287/peerj.preprints.554v2>
 26. Lasken RS, McLean JS (2014) Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat Rev Genet* 15:577–584
 27. Rinke C, Lee J, Nath N, Goudeau D, Thompson B, Poulton N, Dmitrieff E, Malmstrom R, Stepanauskas R, Woyke T (2014) Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat Protoc* 9:1038–1048
 28. Stepanauskas R, Fergusson EA, Brown J, Poulton NJ, Tupper B, Labonté JM, Becraft ED, Brown JM, Pachiadaki MG, Povelaitis T, Thompson BP, Mascena CJ, Bellows WK, Lubys A (2017) Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nat Commun* 8:84. <https://doi.org/10.1038/s41467-017-00128-z>
 29. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Pribelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477
 30. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
 31. Markowitz VM, Chen I-MA, Chu K, Szeto E, Palaniappan K, Pillay M, Ratner A, Huang J, Pagani I, Tringe S, Huntemann M, Billis K, Varghese N, Tennessen K, Mavromatis K, Pati A, Ivanova NN, Kyrpides NC (2014) IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res* 42:D568–D573

SNP Discovery from Single and Multiplex Genome Assemblies of Non-model Organisms

Phillip A. Morin, Andrew D. Foote, Christopher M. Hill,
Benoit Simon-Bouhet, Aimee R. Lang, and Marie Louis

Abstract

Population genetic studies of non-model organisms often rely on initial ascertainment of genetic markers from a single individual or a small pool of individuals. This initial screening has been a significant barrier to beginning population studies on non-model organisms (Aitken et al., *Mol Ecol* 13:1423–1431, 2004; Morin et al., *Trends Ecol Evol* 19:208–216, 2004). As genomic data become increasingly available for non-model species, SNP ascertainment from across the genome can be performed directly from published genome contigs and short-read archive data. Alternatively, low to medium genome coverage from shotgun NGS library sequencing of single or pooled samples, or from reduced-representation libraries (e.g., capture enrichment; *see Ref.* “Hancock-Hanser et al., *Mol Ecol Resour* 13:254–268, 2013”) can produce sufficient new data for SNP discovery with limited investment. We describe protocols for assembly of short read data to reference or related species genome contig sequences, followed by SNP discovery and filtering to obtain an optimal set of SNPs for population genotyping using a variety of downstream high-throughput genotyping methods.

Key words Bioinformatics, Short read archive, Reference-guided assembly, Single-nucleotide polymorphism, Non-model organisms

1 Introduction

Despite the increasing ease, decreasing costs, and a wide variety of methods for sequencing complete genomes [1], there is a continuing need for finding and genotyping specified polymorphic sites, for projects ranging from whole genome association studies [2] to population genetics (e.g., [3] and references therein) and phylogenetics [4, 5] to the identification of functional variants [6, 7]. Once appropriate SNPs are identified, genotypes from hundreds to thousands of individuals can be obtained through cost-effective directed genotyping methods that can make use of small amounts and/or degraded DNA [3, 8].

Prior to highly parallel “next-generation” sequencing (NGS) methods, Sanger sequencing of individual loci was a limiting step, especially when genomic data were not available for PCR primer design [9, 10]. With NGS technologies, generation of targeted-locus [11–13] or whole-genome sequence data is now relatively routine, yielding large amounts of genomic data that can be screened for sites that are variable in a single genome [14] or among individuals [11, 15, 16]. A widely used method of both detecting novel SNPs and obtaining genotypes directly from reduced-representation genomic libraries is the RADseq method [17]. This method typically uses sample sets for detecting thousands to tens of thousands of genomic SNPs from larger numbers of individuals (where high-quality DNA is available from population samples, e.g., *see* Ref. 18). Software and methods for SNP detection and GBS have been previously described (e.g., [19, 20]), and so we will not discuss these methods further.

In this chapter, we will focus on the bioinformatics necessary to screen for SNPs from single-individual genome sequences representative of a species of interest, and from low-coverage genome sequences from one or more individuals and a reference genome of the same or a related species. There are several large consortium genome sequencing projects producing genomic data at an ever increasing rate (e.g., [21, 22]), providing raw data for SNP discovery. Even when a complete, high-coverage genome assembly for a species of interest is not available, generating low-coverage genome data and aligning to a related species can provide relatively quick and inexpensive access to thousands of SNPs in a target species.

The methods in this chapter assume availability of a Unix or Linux computing system with sufficient capacity (memory and storage) to handle Gigabyte datasets. The primary software packages are all freely available, though some additional commercial software is recommended (but not necessary) for some analyses.

2 Materials

The methods in this chapter are based on the versions of software listed in Table 1. There is no one public software package that does all of the steps described below, so we have used multiple freely available programs for data processing. We also describe several scripts that have been written using different programs (e.g., Python, R, Perl), so the programs to run these scripts are necessary. Software can change, so it is important to use the specified versions, or test newer versions to make sure that the commands are performing the same functions and options remain as stated. The best way to do this (and familiarize yourself with the software) is to read the help documents or guidelines for each software package and the specified commands.

Table 1
Required and optional software

Software	Version	Reference or URL
AdapterRemoval	2.0	[23]
ANGSD	0.911	[24]
BCFtools	0.1.19	[25]
BEDtools	2.25.0	https://github.com/arq5x/bedtools2/releases
BWA	0.7.5a	[26]
BLAST+ (optional)	2.2.28+	https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download
Fastq-splitter (optional)	0.1.2	http://kirill-kryukov.com/study/tools/fastq-splitter/ (Perl script)
GATK	2.5–2	[27, 28]
JAVA	1.8.0	https://Java.com
Picard tools	1.92	http://broadinstitute.github.io/picard/
PYTHON	2.6.6	https://www.Python.org
R	3.2.5	[29]
RepeatMasker (optional)	4.0.6	http://www.repeatmasker.org/RMDownload.html
SAMtools	1.2	[30]
SEQTK	1.1-r92	https://github.com/lh3/seqtk
SRAToolkit	2.5.7	http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software/

The version numbers were tested. Later versions should be compatible, but sometimes changes are made that may not be compatible with the scripts and commands as we have presented them

The current versions of the scripts described below are available through GitHub at <https://github.com/PAMorin/SNPdiscovery>. Unless otherwise stated, software is implemented in the Red Hat Enterprise 6 (64 bit) Linux Server operating system. Commands and software should function equally on Unix or Linux, but have not been tested on both.

3 Methods

This chapter assumes that short-read sequence data (typically 50–150 bp, though this includes reads up to approximately 500 bp are produced by some sequencing platforms) are available in one of the forms below, and that the publicly available data have been quality checked to remove adapter sequences and poor-quality reads.

However, we outline adapter trimming steps for raw unprocessed sequence data in Subheading 3.6 for shotgun sequence data (random sequence fragments from genomic DNA libraries) generated de novo for SNP discovery. For our purposes, we will consider two types of data: Publicly available genome assembly data that has already been assembled into a draft genome, and shotgun sequence data. For the latter, the assembled genome scaffold for a related species can be used to assist assembly, or in the absence of a reference genome, de novo assembly can be used to generate contigs to serve as the reference, though in most cases this will result in shorter contigs and potentially lower quality assemblies [31].

All Linux command lines are shown in Courier font, and are prefaced by a comment line with “#” at the beginning to describe the function of the subsequent command. This is to facilitate copy/paste of the commands into a text document or directly into the Linux interface.

3.1 Publicly Available Genome Assembly and Short Read Archive (SRA) Data

The benefits of a reference genome assembly include longer contigs (or scaffolds), quality-checked assemblies, repeat masking, and in some cases, annotation and chromosome assignment. These can all facilitate high-quality SNP discovery and selection based on additional knowledge of genome position and association with known genes or chromosomes. The NCBI web site is a good place to search for reference genomes (<http://www.ncbi.nlm.nih.gov/genome/browse/>), but the organization can be confusing and there are important issues to be aware of. As an example, browsing for the genus “*Orcinus*” results in one record for the genome of *Orcinus orca*, the killer whale. This is a member of the family Delphinidae and could be used as the reference for other species in the family, and even in other families of cetaceans, though potentially at the loss of coverage and some bias in genome region coverage due to divergence. The Genome overview page for *Orcinus orca* (www.ncbi.nlm.nih.gov/genome/14034) summarizes the project information, publications, summaries, and links to the mitochondrial and nuclear genome information (under “Replicon Region”), and Genome Region, with links to the assembled scaffolds. We recommend starting under the Summary, and reviewing the BioProjects links to review the projects contributing to the killer whale genome. These may include multiple biosamples (different animals) and data types. Every project is different, so these need to be reviewed to determine which project(s) has the appropriate assembled contigs/scaffolds and SRA data for re-assembly and SNP discovery. For this example there are two BioProject IDs, and the project data for them are shown in Fig. 1a and b.

Both BioProjects are based on the same, single biosample, and link to the same assembly details, but one links to the SRA

a**Project Data:**

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (total)	30267
Genomic DNA	1
Transcript	28599
Protein Sequences	27870
PUBLICATIONS	
PubMed	3
PMC	3
OTHER DATASETS	
BioSample	1
Assembly	1

▼ Assembly details:

Assembly	Level	WGS	Chrs	Taxonomy
GCF_000331955.2	Scaffold	ANOL00000000	1	Orcinus orca (killer whale)

b**Project Data:**

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (total)	856
WGS master	1
Genomic DNA	1
SRA Experiments	7
Protein Sequences	13
PUBLICATIONS	
PubMed	2
PMC	2
OTHER DATASETS	
BioSample	1
Assembly	1

▼ Assembly details:

Assembly	Level	WGS	Chrs	Taxonomy
GCA_000331955.2	Scaffold	ANOL00000000	1	Orcinus orca (killer whale)

Fig. 1 Screen captures from GenBank genome projects web page for the killer whale genome (*Orcinus orca*; <http://www.ncbi.nlm.nih.gov/genome/14034>). Project data table for two BioProjects associated with this genome, (a) PRJNA189949 and (b) PRJNA167475

experiments (the sequence read data). If there are multiple biosamples, it is important to determine which SRA files are from each individual if the goal is to use either pooled sample data or just a single individual for SNP discovery.

On the BioProject description web page, the Assembly link leads to the project summary. All of the assembly files can be obtained via ftp through the NCBI site: <ftp://ftp.ncbi.nlm.nih.gov/genomes/>. After connecting as a guest, you can navigate to the species directory and view the README file that describes each of the directories and files within. For our purposes, the important files are in the directory CHR_Un, which contains the “Chromosome directories,” or “concatenated sequence data for scaffolds that have been assembled from individual GenBank records,” in a variety of formats. The files ending in “.fa.gz” are contigs in compressed FASTA format, and the ones ending in “.mfa.gz” are repeat-masked contigs in compressed FASTA format. These can be downloaded by a variety of methods. Probably the most useful in a Linux environment is to use “wget,” for example:

```
# transfer reference genome file (compressed
# FASTA)
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/Orcinus_orca/
CHR_Un/oor_ref_Oorc_1.1_chrUn.mfa.gz

# decompress the file.
gunzip oor_ref_Oorc_1.1_chrUn.mfa.gz

# Index your reference using the faidx function
# in SAMtools to allow efficient random access
# to the reference bases.
# This creates a file with the same name, but
# ending with “.fai”.
samtools faidx oor_ref_Oorc_1.1_chrUn.mfa

# Create index files needed for BWA read
# alignment to the reference.
# This creates 5 files with the same file name
# but different suffixes.
bwa index oor_ref_Oorc_1.1_chrUn.mfa
```

File transfer may take several hours, depending on the size of the file and speed of connection. It is usually not advised to map reads to a repeat masked genome, as reads neighboring the masked repeats may not map, thereby reducing coverage around the repeats. However, if you have a reference genome with reasonably high coverage, the loss will be minimal relative to the amount of high-coverage genome data, and using the repeat-masked reference sequence will reduce the need for filtering out SNPs in repeats later.

The next step is to identify and download the short-read archive data.

Clicking on the SRA link in the “Project Data” table on the BioProject page (e.g., “7” under “Number of Links” in the killer whale example) produced a list of the SRA experiments, each of which may have ≥ 1 sequence run files. You can look through these on the web site to identify the type of sequence (e.g., Illumina

HiSeq 2000, paired-end) and the size of the individual files. Once you know which files you want to use, you can download individual files from the SRA portal <http://sra.dnanexus.com>. At the site, you can search for your species, select the appropriate study (if there is more than one), and click the “related runs” button to show all of the sequencing runs associated with that project. For this example, the project with accession SRP015826 has seven experiments with 16 runs total, representing the 16 short-read archive files. The files can be selected by checking the boxes of those you want and using the “Download” button. This generates a web page with the download buttons for individual files, or you can select files and click the button for “download SRA URLs” to download a text file containing all of the URL links to facilitate command-line transfer of the files using the “wget” command in a Linux environment:

Example SRA URLs file: `download_sra_urls.txt` (for the *Orcinus orca* project):

```
ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-
instant/reads/ByRun/sra/SRR/SRR574/SRR574967/
SRR574967.sra
ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/
reads/ByRun/sra/SRR/SRR574/SRR574968/SRR574968.sra
ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/
reads/ByRun/sra/SRR/SRR574/SRR574969/SRR574969.sra
(etc.)

# Transfer each of the files using “wget” and
# the URLs in the download_sra_urls.txt
# document:
wget ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/
reads/ByRun/sra/SRR/SRR574/SRR574967/SRR574967.sra
```

These files are often very large, and the combined files can be several hundred GB, so the transfer may take a long time and require a lot of storage space on your system. It is important to determine how much storage space will be needed prior to starting to be sure you have sufficient storage capacity for the raw files and the assemblies. Although this is difficult to estimate and depends on whether or not you delete file that are no longer needed (e.g., delete SAM files after converting to BAM format), you should expect to use about 20 times as much disk space as the size of your decompressed short-read archive fastq file(s).

SRA files are compressed binary files and need to be expanded into FASTQ files prior to use in an assembly. This requires the `fastq-dump` function from the SRAtoolkit software. Files size will expand substantially, typically by 4–5× the original SRA file size.

```
# Convert SRA files into FASTQ files
fastq-dump SRR574967.sra --split-3
```

The “--split-3” option indicates that the file contains paired-end reads and will be split into two fastq files, one for read 1 and a second for read 2 (e.g., in this case, SRR574967_1.fastq, SRR574967_2.fastq). If the file only contains single-read data, only a single file will be generated, so it is recommended that this option is always used to ensure data are correctly parsed into files.

The FASTQ file may be too large to be processed when only one SRA file is available on the SRA portal for a high coverage genome. Fastq-splitter can be used to divide the large file into several smaller files that are easier to handle. After downloading the program, you need to make sure that the perl script is set as an executable.

```
# Command to check if the perl script is set to
# be executable
chmod +x fastq-splitter.pl
# Split the file in 6 parts for both the
# forward and reverse reads
./fastq-splitter.pl --n-parts 6 --check
SRAfileF.fastq

./fastq-splitter.pl --n-parts 6 --check
SRAfileR.fastq
# The path to the perl script (fastq-
# splitter.pl) should be given.
# --n-parts indicates into how many files the
# large
# SRA file should be divided (6 in this
# example).
# --check does some verification of FASTQ
# format correctness.
```

3.2 Creating a De Novo Set of Reference Contigs

A draft genome sequence is not necessary to align sequence reads for SNP discovery, but some reference is necessary. In the absence of a reference genome, a set of reference contigs can be created by de novo assembly of short-read data. The details of de novo assembly are not the focus of this chapter, but there are commercially available programs (e.g., CLC Genomics Workbench (CLCbio), Geneious (BioMatters Limited)) and freely available programs such as SOAPdenovo2 [32], DISCOVAR de novo (Broad Institute), SGA [33], some of which are optimized for different operating systems or types of data (e.g., see list at <http://omictools.com/genome-assembly-category>). The goal of using these packages is to generate high-quality contigs, then use them as the reference sequences to re-align reads for SNP discovery. Once a set of de novo assembled contigs has been produced (in FASTA format), they can be treated the same as a reference genome sequence for read alignment and SNP discovery (*see* Ref. 31).

3.3 Alignment of Reads to Reference Contigs

Alignment of reads to reference contigs is fairly straightforward, but requires multiple steps that can be confusing. There are multiple alignment programs, but one of the most frequently used and easiest to use is the Burrows-Wheeler Aligner (BWA; [26]).

When there are multiple smaller FASTQ files for a single individual, they can be concatenated into a single file for each read direction. This is a Linux function:

```
# Concatenate forward read files.
cat filename1_read1.fastq filename2_read1.fastq >
forward_reads.fastq
```

However, when the files are large (e.g., >10 GB), running separate alignments of each FASTQ file (or pair of files for paired-end data) will improve speed by allowing parallel processes run on separate cores (*see Note 1*). The resulting alignments can be combined after conversion to the smaller BAM files (*see below*).

Although there are many options that can be altered to vary alignment parameters, the updated aligner “BWA-MEM” algorithm [34] is implemented in BWA and has been optimized for aligning short reads to a large genome such as human without the need to specify alignment parameters, and performs as well as or better than many other aligners with short read (100 bp) and longer read-length data [34].

Mitochondrial DNA can make up a significant portion of sequence reads, so removal of reads that map to the mitogenome can reduce file sizes and speed up subsequent analyses. A reference mitogenome FASTA file can typically be downloaded from GenBank for the target species or a closely related species, or the mitogenome sequence can often be assembled *de novo* from the NGS reads (e.g., [35]). Map the reads to the mitogenome with BWA, convert the SAM file to a BAM file, and sort with SAMtools.

```
# Align reads (reads.fq) to indexed reference
# mitogenome # (ref_mtgenome.fasta)
bwa index ref_mtgenome.fasta

bwa mem ref_mtgenome.fasta reads.fq >
mapped_to_mt.sam
# Note, if there are paired-end reads, the two
# respective files are listed, e.g., reads_1.fq
# reads_2.fq.

# Convert the SAM file to a sorted BAM file.
samtools view -b -h mapped_to_mt.sam | samtools
sort - mapped_to_mt_sorted
# output = "mapped_to_mt_sorted.bam"
```

The BAM file will also retain the reads that did not map to the mitogenome reference, i.e., the reads from the nuclear genome that we are interested in. These unmapped reads can be extracted to a new BAM file, and then converted to a FASTQ file.

```
# Copy reads that did not map to the mitogenome
# to a new file and convert from BAM file to
# FASTQ format.
```

```
samtools view -f4 -h mapped_to_mt_sorted.bam >
unmapped.bam
```

```
samtools bam2fq unmapped.bam > reads1.fq
# For paired-end reads:
cat reads.fq | grep '^@.*/1$' -A 3 --no-group-
separator > reads1.fastq
cat reads.fq | grep '^@.*/2$' -A 3 --no-group-
separator > reads2.fastq
```

Align the unmapped (non-mtDNA) reads to the reference genome using BWA. BWA requires an indexed FASTA reference file (see “bwa index” above). The output file is in SAM format.

```
# Map reads (reads1.fq) to the reference FASTA
# file (ref.fa):
bwa mem -t2 ref.fa reads1.fq > aln_sel.sam
# -t option = number of threads to use
# (default = 1).
```

The output SAM files are typically very large, and can be reduced to the binary form, or BAM file format, for subsequent analyses. The SAM files can then be deleted, as they are no longer needed.

```
#Convert SAM file to BAM file and sort BAM
# file.
samtools view -@1 -T ref.fa -b aln_sel.sam |
samtools -@1 sort - aln_sel_sorted
# -@ indicates the number of threads to use.
# Default = 1. This option is only available in
# SAMtools for the 'view' and 'sort' functions.
# -b Output in the BAM format
# -T FASTA format reference file.
```

PCR duplicates should be removed using the rmdup function of SAMtools. This is very important, as PCR amplification can bias the distribution of reads across loci and alleles.

```
# Collapse clonal reads
samtools rmdup -S aln_sel_sorted.bam aln-
sel_sorted_unique.bam
# NOTE: There is a bug in Samtools 1.2 that
# causes this to not work. Both earlier and
# later (v1.3) versions have been reported to
# work.
```

If there are several BAM files *for an individual* (e.g., aln_sel_sorted_unique.bam, aln-se2_sorted_unique.bam, etc.), they need to be merged into one file (individual1.bam).

```
# Merge the files
samtools merge -@2 individual1.bam
aln_sel_sorted_unique.bam
aln_se2_sorted_unique.bam
aln_se3_sorted_unique.bam
aln_se4_sorted_unique.bam
# -@2 indicates that 2 threads are used.
```

At this stage, the sorted BAM file contains both mapped and unmapped sequencing reads. You may therefore want to remove the unmapped sequencing reads to save on disk storage space and improve the efficiency of your downstream analysis. But beware that your unmapped reads may contain useful data, such as reads associated with the Y-chromosome if you mapped your reads to a female reference, or microbiome data, etc.

```
# Keep reads that have mapped to the reference
genome with a qscore ≥30.
samtools view -bh -F4 -q 30 individual1.bam >
individual1_q30.bam
```

If the sequence reads have been aligned to a related species rather than reference contigs from the same species, a new target-species reference needs to be generated from the merged alignment file, and the alignment process is started over. This results in the reads being mapped to contigs of the same species so that inferred SNPs are due to intra-specific variation rather than differences between the reference and target species sequences.

```
# Extract the consensus sequences of the
# target-species contigs from the alignment.
samtools mpileup -uf ref.fa individual1.bam |
bcftools call -c | vcfutils.pl vcf2fq >
target_consensus.fq
# Convert the output fastq file to a fasta file
# using SEQTK to use as the new reference
# sequence, and index using 'bwa index' as
# above.
seqtk seq -a target_consensus.fq >
target_consensus.fasta
```

3.4 Pipeline 1: SNP Discovery from High-Coverage Genome Assembly

SNP discovery from alignment files (BAM) with medium to high average coverage ($>10\times$) uses depth and quality criteria, and extracts SNPs with flanking sequence data for assay design (e.g., for AmpliSeq, GTseq, TaqMan, or capture-enrichment GBS). We describe a simple SNP discovery pipeline that uses the variant finder function of BCFtools [25] to scan an alignment (BAM file) for SNPs that meet specified criteria. The minimum and maximum coverage criteria are dependent on the average coverage of the genome alignment. When coverage is moderate (i.e., $10\text{--}30\times$), we set the minimum depth of coverage for a SNP at 10 reads, and a maximum depth of 100 in order to ignore potentially collapsed duplicated sequence in the assembly. However, when the average depth of an alignment is high (i.e., $>30\times$), the criteria should be changed to reflect the range from $1/3$ to twice the average coverage (see “Estimate coverage” in Subheading 3.8).

The BAM file to be used for SNP discovery needs to be indexed for SAMtools.

```
# Index BAM file for SAMtools.
samtools index individual11_q30.bam
```

The SNP discovery pipeline is managed through a shell script (*see* Appendix 1) that performs several functions. The shell script is launched by executing the script in a directory that also must contain the reference FASTA file (previously indexed with samtools faidx), the alignment BAM file (sorted) and the SNP discovery Python script “Script2_generate_genotype_blocks.py” (*see* Appendix 2). The shell script takes as input the alignment BAM file (sorted), reference FASTA file, and the total desired bases of the flanking sequence of the SNP (by default 300 bp). The shell script takes objects in order, assigning them to aliases that are used to fill in the appropriate filenames into a series of commands. Execution of the shell script first runs samtools mpileup to compile the read bases, quality, and coverage information at each site in the alignment, then calls SNPs from the pileup using BCFtools, using filters to exclude loci with depth of coverage below and above specified levels (set within the shell script; *see* Appendix 1). The custom Python script “Script2_generate_genotype_blocks.py” further filters the SNPs based on user-defined quality cutoff and flanking sequence length parameters, and outputs two results files (*see* Note 2).

```
# Execute shell script1 for reference FASTA
# file and alignment BAM file.
./Script1_SNP_call_filter.sh
individual11_q30.bam ref.fa 300 output
# '300' indicates the sequence length,
# specifying 150bp on either side of the target
# SNP.
# 'output' represents the file name that will
# be assigned to the output files.
```

The output files from this script are a FASTA file containing the designated target sequences surrounding each SNP, and a genotype VCF file. The VCF file format is described in detail elsewhere (<https://github.com/samtools/hts-specs>). This file contains information about each SNP, starting with the contig ID (“chrom”), the SNP position, the reference and alternate alleles, a quality score, and a list of additional “info” about the SNP, including the depth of coverage (DP), various quality measures, the count of forward and reverse reads for each allele used in variant calling (DP4), and the consensus quality (FQ) (*see* Fig. 2). The total number of reads in DP4 may be lower than DP because low-quality bases are not counted. Given one sample in the alignment, the consensus quality (FQ) is typically positive for heterozygous loci, and negative for homozygous loci (*see* more details at <http://>

#CHROM	POS	I	REF	ALT	QUAL	FIL	INFO (DP=depth,	FORMAT
		D				TER		
z0000161_c	270	.	C	T	50.0072	.	DP=10;VDB=0.748679;SGB=-	GT:PL
ontig_4391							0.511536;RPB=0.809011;MQB=0.924584;MQSB=0.974597;BQB=0.924584;MQOF=0	
0							;AF1=0.5;AC1=1;DP4=3,3,2,1;MQ=60;FQ=53.0153;PV4=1,0.249156,1,1	
z0000161_c	371	.	G	A	56.0072	.	DP=10;VDB=0.803107;SGB=-	GT:PL
ontig_4391							0.556411;RPB=0.89338;MQB=0.974597;MQSB=0.924584;BQB=0.730948;MQOF=0;	
7							AF1=0.5;AC1=1;DP4=2,3,1,3;MQ=60;FQ=59.0154;PV4=1,0.224349,1,1	
z0000161_c	1694	.	A	G	104.008	.	DP=10;VDB=0.381141;SGB=-	GT:PL
ontig_4396							0.616816;RPB=0.783809;MQB=1.00775;MQSB=1.00775;BQB=0.895781;MQOF=0;A	
4							F1=0.5;AC1=1;DP4=2,2,2,4;MQ=60;FQ=79.0088;PV4=1,0.278763,1,0.237204	
z0000161_c	358	.	C	T	87.0076	.	DP=10;VDB=0.277213;SGB=-	GT:PL
ontig_4425							0.590765;RPB=0.487298;MQB=0.974597;MQSB=0.924584;BQB=0.406082;MQOF=0	
2							;AF1=0.5;AC1=1;DP4=2,2,4,1;MQ=60;FQ=89.9475;PV4=0.52381,0.0880959,1,0.0674986	
z0000161_c	1951	.	T	C	77.0075	.	DP=10;VDB=0.66971;SGB=-	GT:PL
ontig_4442							0.556411;RPB=1.00775;MQB=1.00775;MQSB=0.916482;BQB=0.783809;MQOF=0;A	
1							F1=0.5;AC1=1;DP4=4,2,3,1;MQ=60;FQ=80.0156;PV4=1,0.304798,1,1	
z0000161_c	362	.	G	A	108.008	.	DP=14;VDB=0.651328;SGB=-	GT:PL
ontig_4493							0.616816;RPB=0.991701;MQB=1;MQSB=1;BQB=0.991701;MQOF=0;AF1=0.5;AC1=1	
9							;DP4=6,2,5,1;MQ=60;FQ=111.016;PV4=1,0.470145,1,1	

Fig. 2 Example of Variant Call Format (VCF) file format. Column headings are CHROM (contig ID), POS (SNP position), ID (sample IDs for some data types), REF (reference allele), ALT (alternate allele), QUAL (quality score), FILTER (filters that have been applied to the data), INFO (combined information types), FORMAT (genotype format codes), and additional genotype or sample file information. Additional details can be found at <https://samtools.github.io/hts-specs/VCFv4.1.pdf>

samtools.sourceforge.net/mpileup.shtml). Although it seems counterintuitive that there would be homozygous SNPs, they may be frequent in some datasets, especially if the reference sequence is different from the reads mapped to it (different species, subspecies, or even just a different individual fixed for the alternate allele), and when reads are mapped incorrectly to the reference. A negative FQ value may also indicate that one of the alleles is rare.

The overall depth and the count of each base at the SNP site can be useful for further filtering of SNPs based on the number of nucleotides at the SNP position and their frequency. For example, if the SNPs are from a single individual, you would expect approximately even distribution of two alleles at heterozygous SNPs. A SNP with DP = 20 and DP4 = 0,0,11,0 (No. of forward ref. alleles, reverse ref. alleles, forward non-ref. alleles, reverse non-ref. alleles) would indicate that even though there is a total depth of 20, only 11 reads met the quality threshold, and all of them were forward reads of the non-reference allele, so this is not likely to be a high-quality candidate SNP. A SNP with DP = 20 and DP4 = 7,0,13,0 would have close to the expected 50% each of two alleles and all 20 reads passed the specified quality filters.

3.5 Pipeline 2: SNP Discovery from Low-Coverage Multiplex Shotgun Data

SNPs can be called from low-coverage shotgun sequencing data (<1x) from multiple samples using genotype likelihoods, therefore taking uncertainty in base-calling into account [36]. This section describes read sequence data processing, genotype likelihoods

estimation, and a filtering pipeline for SNP discovery using this approach (see [37]).

Data requirement: FASTQ formatted sequence read data from multiple samples. As an example, one lane of Illumina HiSeq or NextSeq sequencing of around 30 individuals (genome size of 2.5 Gb) would be suitable for genome-wide SNP discovery [37]. In the latter study, mean sequencing read depth per site considering all the samples together was approximately 5 \times , but the majority of the sites had sequencing reads from just a small proportion of the individuals sampled, each sequenced at <1 \times coverage. We describe data processing for de novo generated read data (i.e., generated by the investigator specifically for SNP discovery) and unmasked reference sequences, followed by SNP calling that can be applied to either the assembled de novo sequence data or SRA data assembled to a masked reference genome from GenBank.

3.6 Sequence Data Processing

This section assumes that single- or paired-end read data have been de-multiplexed to separate reads by the index sequences of each sample in the multiplex pool. SRA files downloaded from GenBank have presumably been processed to remove adapters and filtered for quality. For data generated specifically for SNP discovery, these steps need to be done prior to mapping reads to the reference genome. For each FASTQ file, ADAPTER-REMOVAL (or another program, e.g., Trimmomatic; [38]) can be used to remove adapter sequences from the sequencing reads and remove sequence reads that are ≤ 30 bp or another length threshold following trimming.

```
# Remove adapter sequences.
AdapterRemoval --file1 ind1.fastq --basename
indltrimmed --minlength 30 --trimns --
trimqualities --minquality --gzip
# ind1.fastq is the input file and indltrimmed
# the output file.
# --trimns: Remove stretches of Ns from the
# sequencing reads in the 5' and 3' end.
# --trimqualities: consecutive stretches of low
# quality bases (the quality threshold is set
# by minquality) are removed from the 5' and 3'
# end of the reads. All bases with minquality
# or lower are trimmed. If 'trimns' is also
# indicated, stretches of low-quality bases
# and/or Ns are trimmed.
# --minquality: quality threshold for the
# trimming of low quality bases. Default is 2.
# --minlength 30: reads that are > 30 bp are
# kept.
# --gzip: the output file is compressed. There
# will be 2 output files: one with the
# discarded sequencing reads
# (indltrimmed.discarded.gz) and one with the
# retained reads (indltrimmed.truncated.gz).
# --qualitybase: specifies the platform-
```

```
# specific base quality score encoding, with
# the assumed default of Phred+33 used by
# Illumina. AdaptorRemoval can also handle
# Phred+64 and 'Solexa'encoded quality scores,
# but this input should be specified using the
# --qualitybase command option.
```

Reads in FASTQ format must then be mapped to the reference genome of the studied species or a related species, then converted to a sorted BAM file following the same steps outlined above. The reference can be repeat hard-masked, or repeats can be masked in the BAM file following mapping.

To hard mask the reference FASTA file to which reads are being mapped, we use RepeatMasker. Repeat libraries must also be accessed from the RepBase database which is available through the GIRI Web site (<http://www.girinst.org/rebase/index.html>). Note that this requires opening an account, however, there is currently no charge for this for academic non-profit users.

```
# Mask repeats
RepeatMasker genome.fa -nolow -norna -specie
Cetartiodactyla
# genome.fa is the input reference fasta file.
# -nolow specifies that only interspersed
# repeats are masked and STRs and low
# complexity regions are retained.
# -norna prevents the default masking of small
# RNAs
# -specie specifies the taxonomic group the
# query sequence belongs to, in our examples
# the Cetartiodactyla, and then checks the
# RepBase file for known repetitive elements
# for this taxonomic group.
```

The output files include a masked copy of the input file, which contains the query sequences, with identified repeats and low complexity sequences masked and replaced with “N”s. A map file is also generated, in which coordinates in terms of contig, start and end positions are given, which can then be used as a .bed file (*see* below). The advantage of this is that reads can be mapped to the unmasked reference and then removed, which increases coverage in regions adjacent to repetitive elements.

```
# Run BEDtools to remove repeated regions.
intersectBed -abam data.bam -b RepeatMask.bed -
v | samtools view -h - | samtools view -bSh - >
data_noRepeats.bam
# The BAM file from which the repeated regions
# should be removed is given with the option
# -abam and the file containing the repeated
# regions with -b.
```

Follow the steps described above to download a reference genome file in FASTA format, and index it for processing with

SAMtools and BWA. Align and remove mitogenome reads, align remaining reads to the reference, convert the resulting SAM file to BAM format, collapse clonal reads, merge multiple files for single samples (if necessary), and remove poor-quality alignments ($\text{MAPQ} < 30$) as previously described (*see* Subheading 3.3 above).

A “bamlist” file of the path to the BAM files for each individual needs to be created. It is a text file (data.bamlist) with a line for the path to each individual BAM file. If they are in the same directory as the ongoing analysis, then just the file names are listed.

```
individual1_q30.bam
individual2_q30.bam
# etc.
# NOTE: if the files are not in the current
# directory, the path must be specified (e.g.,
# /path/individual1_q30.bam, etc.). See
# documentation for ANGSD for more details.
```

3.7 SNP Calling Using Genotype Likelihoods for Low- Coverage Alignments

SNP calling is performed using genotype likelihoods that take uncertainty into account and can be estimated in ANGSD [24] from multiple samples as recommended by Nielsen et al. [36]. Uncertainty in SNP discovery in low coverage data can be the result of sequencing, base-calling, mapping, and alignment errors. Briefly, genotype likelihoods are calculated using probabilistic methods and take the quality scores of the sequencing reads into account [36]. A base is considered a SNP if the minor allele frequency is significantly different from 0 as inferred from a likelihood ratio test [39]. The threshold is set with the “-SNP_pval” option and $P < 0.000001$ is considered as conservative. The genotype likelihood is calculated using the SAMtools method with the “-GL1” option [24, 25] to infer the major and minor alleles with `-doMajorMinor1` and to estimate major and minor allele frequencies using the EM algorithm with `-doMaf 2` [39, 40] (*see Note 3*).

```
# Infer SNPs from the alignments listed in
# "data.bamlist" using genotype likelihoods in
# ANGSD.
angsd -GL 1 -out genolike_data -nThreads 2 -
doGlf 2 -doMajorMinor 1 -SNP_pval 1e-6 -doMaf 2
-minQ 30 -bam data.bamlist
# -out indicates the output file (followed by
# the filename)
# -bam indicates the input file (followed by
# the file name), which is the data.bamlist
# file created previously.
# -doGlf 2 creates a genotype likelihood beagle
# file.
# -nThreads indicates the number of threads.
# minQ 30 to set the minimum base quality
# (minQ) to 30
# Optional settings: -minMapQ 30 to set the
# minimum mapping quality (minMapQ) to 30 (if
# not filtered before).
```


Running ANGSD produces a compressed .mafs output file: `genolike_data.mafs.gz`. The file must be decompressed prior to use in the filtering steps (below).

```
#decompress .mafs file
gunzip genolike_data.mafs.gz
```

3.8 SNP Filtering

A filtering pipeline is then applied to further avoid bias linked to next-generation-sequencing and low coverage data. The different filters are:

- *Filter 1: discarding SNPs in regions of poor mapping quality ($MAPQ \leq 30$).*
- *Filters 2 and 3: removing SNPs in regions of excessive coverage and high numbers of individuals:* Filter 2 removes SNPs that are in regions with twice the mean coverage. As this is a rather arbitrary threshold, we advise plotting and exploring the distribution of coverage at this stage. Illumina shotgun sequencing data typically results in a Poisson distribution of different coverage depths, and so selecting the point at which the upper bound of depth of coverage starts to deviate from this distribution can be a more formal approach to set a threshold for determining regions of excessive coverage.
- Filter 3 removes SNPs that are found (i.e. for which there is coverage) in a high number of individuals. The cutoff is defined by the upper tail of the distribution of the plot of the number of SNPs against the number of individuals. SNPs in the upper tail of the distribution are removed. The high coverage of these SNPs may be because they are in unmasked repeated regions, in nuclear mitochondrial DNA inserts (NUMTs), or other mapping artifacts (i.e., paralogous loci). This filter is only recommended with ultra-low coverage data.
- *Filter 4: discarding SNPs that have an estimated minor allele frequency (MAF) ≤ 0.05* as it is the smallest MAF that we could theoretically have with a cutoff of ten individuals (threshold defined earlier but it may change depending on the data analyzed). However, estimations of genotype likelihoods should reduce the error rate. Additionally, rare variants are important for several applications such as the inference of demographic history based on the Site Frequency Spectrum and the estimation of several population genetic parameters [41, 42]. Hence, depending on the downstream analyses it may be worth keeping those SNPs.
- *Filter 5: Select only SNPs with flanking sequence of a specified length* (e.g., 150 bp on either side of the target SNP) and with no N's in the flanking region, to allow design of primers and/or probes.
- *Filter 6: SNP validation.* (Optional) This step is not necessary but if available, the identified SNPs could be compared to already existing genomic resources such as published high cov-

erage genomes to provide further confidence in the discovered SNPs, and if they are from different geographical areas, to identify a set of globally shared variants.

- *Filter 7: (Optional) Remove SNPs with excess heterozygosity.* This step can only be conducted meaningfully when there are sufficient samples to infer a significant deviation from the expectations of Hardy-Weinberg equilibrium (HWE), and is useful in identifying putative SNPs that are the result of gene duplications, gene families, and repetitive regions.
- *Filter 8: (Optional) Compare SNP locus sequences to reference databases (e.g., GenBank) to filter out loci that might match to non-target species, duplicated loci, gene families, and repeat regions.*

The commands to run these eight filtering steps are detailed below.

Filters 1 and 2 (removing SNPs in regions of poor mapping quality (MAPQ \leq 30) and SNPs that are in regions with twice the mean coverage): Commands will be the same for filters 1 (poor mapping quality) and 2 (excessive coverage), the two scripts are provided and the differences between the two filters are highlighted below. The mapping quality filter is only needed if the regions of high mapping quality (MAPQ $>$ 30) were not previously selected using samtools view (*see* Subheading 3.3 and Note 4).

First, coverage is estimated in ANGSD using the doDepth function. This function estimates the distribution of the depth of coverage for each individual as well as across all individuals.

```
# Estimate coverage
angsd -bam data.bamlist -doDepth 1 -out data -
doCounts 1 -nInd 30 -minMapQ 30 -minQ 30
# Reads with a mapping quality (minMapQ) above
# 30 and nucleotide # qscore (minQ) above 30 are
# kept.
# -nInd corresponds to the number of sequenced
# individuals.
# -doDepth 1 function in angsd also requires
# the -doCounts 1 option to estimate coverage.
# -out indicates the output file, followed by
# the output filename.
```

Three output files are produced, including data.depthSample and data.depthGlobal. In data.depthSample each line represents one sample (i.e., one individual) and column 1 is the number of sites with a depth of 0 \times , column 2 the number of sites with a depth of 1 \times , etc. In data.depthGlobal the number of sites for each depth category is given across all samples. The mean depth across all samples can be calculated using this file (*see* Note 5). It should be noted that the sequencing depth for all sites and not only SNPs is provided here.

```
# generate distributions of depth and quality
# scores:
angsd -b data.bamlist -doQsDist 1 -doCounts 1
-maxDepth 100 -doDepth 1 -out bam.qc
# data.bamlist is the same text file list of
# bamfiles as used above.
# -doQsDist 1 counts the number of bases for
# each quality score
# -maxDepth 100 sets the maximum depth at 100;
# sites covered at depth >100 are combined,
# which may produce a peak at 100 for samples
# with higher coverage.
# -out indicates the output file, followed by
# the output file name.
```

This command produces four output files: `bam.qc.arg`, `bam.qc.depthGlobal`, `bam.qc.depthSample`, and `bam.qc.qs`. These files can be plotted using the R script `Script10_plotQC.R` (see Appendix 12). The script creates a pdf file of the plotted output, and the file “`bam.qc.info`” with the q-score and global depth data.

As an example, we consider here that the mean coverage across all samples is $5\times$ and that regions with coverage $>10\times$ are potential unmasked repeated regions or mapping artifacts. Regions of poor mapping quality ($Q < 30$) and excessive coverage ($>10\times$) are detected using the `CALLABLELOCI` tool in GATK. Running `CALLABLELOCI` involves some data formatting.

```
# Create a dictionary file on the initial
# reference using Picard tools.
java -jar path_to/picard.jar
CreateSequenceDictionary R= ref.fa O= ref.dict
# The output file must have the same stem name
# as the reference file so that it can be used
# by GATK based on the reference file name.
```

If using low coverage data it may be best advised to merge individual BAM files into one consensus sequence to better identify putative repetitive regions or other mapping artifacts.

```
# merge individual bam files
samtools merge data.bam individual1_q30.bam
individual2_q30.bam individual3_q30.bam #etc.

# The read groups are added to the header using
# Picard tools.
java -jar -Xmx10g path_to/picard.jar
AddOrReplaceReadGroups INPUT= data.bam OUTPUT=
data_withheader.bam SORT_ORDER=coordinate
RGID=sample RGLB=sample RGPL=illumina
RGSM=sample RGPU=name CREATE_INDEX=true
# -Xmx10g; defines the maximum memory size for
# Java, e.g. 10g = 10GB
# Sort_order: to order the output file (if not
# specified the order in the output file is the
```

```
# same as in the input file)
# RGID: Read Group ID Default value: 1
# RGLB: Read group library required
# RGPL=illumina, can be changed for other NGS
# platforms
# RGSB: Read group sample name required
# RGPU: Read group platform unit (e.g. run
# barcode) required
# CREATE_INDEX: create a BAM index when writing
# a coordinate-sorted BAM file
```

CallableLoci in GATK can then be run to identify the regions of poor mapping quality (MAPQ < 30) and excessive coverage (>2× mean coverage).

```
# Run CallableLoci in GATK.
java -jar path_to/GenomeAnalysisTK.jar -T
CallableLoci -R ref.fa -I data_withheader.bam
--maxDepth 10 -mmq 30 -summary
data_callable.summary -o data_callable.bed
# -T = Name of the tool to run.
# -R = The reference genome against which the
# sequence data was mapped. The GATK requires
# an index file (samtools faidx) and a
# dictionary file accompanying the reference.
# They need to have the same filename, ending
# in .dict for the dictionary file and .fai for
# the indexed file.
# -I = Input file containing sequence data (BAM
# or CRAM).
# --maxDepth corresponds to twice the mean
# coverage (global). In this example the mean
# coverage is 5×.
# -mmq: The minimum mapping quality (filter 1:
# regions with a mapping quality lower than
# this threshold will be considered of poor
# mapping quality).
```

The lines of the data_callable.bed file look like the following (a few example lines have been extracted from different parts of the file):

```
JH472447    0          21          NO_COVERAGE
JH472447    21         65          LOW_COVERAGE
JH472447    65         80          CALLABLE
JH472447    80        133         LOW_COVERAGE
JH472447   133       150         CALLABLE
JH472447   601657    601658     LOW_COVERAGE
JH472447   601658    601711     POOR_MAPPING_QUALITY
JH472447   601711    601712     CALLABLE
JH472447   601712    601719     LOW_COVERAGE
JH472461   671807    671914     CALLABLE
JH472461   671914    671947     NO_COVERAGE
JH472461   671947    671951     LOW_COVERAGE
JH472461   671951    671966     CALLABLE
JH472461   671966    671996     EXCESSIVE_COVERAGE
```

```
JH472461 671996 671997 CALLABLE
JH472461 671997 671999 LOW_COVERAGE
```

This file contains all sites. It indicates the contig, scaffold or chromosome depending on the available information, the start and end positions of the region, and then the characteristic of the region (i.e., whether the region has no coverage, low coverage, poor mapping quality, excessive coverage or if the region is callable). For example on scaffold JH472447 bases from positions 601658 to 601711 are in a region of poor mapping quality and on scaffold JH472461 bases from position 671966 to 671996 are in a region of excessive coverage.

Two similar R scripts are used to select independently the regions with poor mapping quality and excessive coverage.

```
# Source the script (Appendix 3) file using
# "Rscript" or execute individual lines from the
# script in the R environment.
Rscript Script3a_Filter1_mapping_quality.R
# Script is written to use the GATK output file
# named "data_callable.bed".
```

The “poor_mapping_quality.txt” output file looks like the following:

```
JH472447 6307 6341 POOR_MAPPING_QUALITY
JH472447 6437 6475 POOR_MAPPING_QUALITY
JH472447 516203 516207 POOR_MAPPING_QUALITY
JH472447 601658 601711 POOR_MAPPING_QUALITY
```

```
# Source the excessive coverage script
# (Appendix 4) file using "Rscript" or execute
# individual lines from the script in the R
# environment.
Rscript Script3b_Filter2_excessive_coverage.R
# Script is written to use the GATK output file
# named "data_callable.bed".
```

The “excessive_coverage.txt” output file look like the following:

```
JH472456 305636 305662 EXCESSIVE_COVERAGE
JH472461 129045 129049 EXCESSIVE_COVERAGE
JH472461 671966 671996 EXCESSIVE_COVERAGE
JH472465 237841 237878 EXCESSIVE_COVERAGE
```

Then, the SNPs that are in regions of poor mapping quality are removed from the `genolike_data.mafs` file (from the Genotype Calling section above) using a second R script (*see* Script 4a below) and the “poor_mapping_quality.txt” filter.

The `genolike_data.mafs` (from Subheading 3.7) file looks like the following:

```
chromo position major minor unknownEM pu-EM nInd
JH472447 724 G C 0.406400 2.160509e-07 5
```

JH472447	6226	C	A	0.378201	5.862932e-11	8
JH472447	599458	G	A	0.432863	1.945000e-12	4
JH472447	601678	A	T	0.301786	4.667339e-10	11

In the first column (i.e., “chromo”), the contig, scaffold, or chromosome number is given (depending on the assembly level of the reference genome). The second column corresponds to the position on the scaffold, contig, or chromosome. The major and minor alleles are given. “unknownEM” corresponds to the minor allele frequency. “pu-EM” is the p -value indicating that the minor allele frequency is significantly different from 0 (i.e., that the base is a SNP). “nInd” indicates the number of individuals for which there is coverage at each SNP.

The SNP at position 601678 on scaffold JH472447 in the .mafs file is in a region of poor mapping quality (JH472447, positions 601658–601711). The R script is used to filter this SNP out. All the other SNPs in these example lines are not in a region of poor mapping in the example lines of the “poor_mapping_quality.txt” filter given above and will be kept.

```
# Source the "remove_poor_mapping_quality"
# script (Appendix 5) file using "Rscript" or
# execute individual lines from the script in
# the R environment.
Rscript
Script4a_Filter1_Remove_poor_mapping_quality.R
# This script requires the input files
# "genolike_data.mafs" and
# "poor_mapping_quality.txt".
# The output file is "SNPs_goodquality.txt"
```

A similar script is used to remove the SNPs in regions of excessive coverage. The SNPs in those regions are likely to be in a repeat region, NUMT or in a region with some other mapping artifact. The script is run on the output file of Filter 1 (SNPs_goodquality.txt) using the “excessive_coverage.txt” filter.

```
# Source the "remove_excessive_coverage" script
# (Appendix 6) file using "Rscript" or execute
# individual lines from the script in the R
# environment.
Rscript Script4b_Filter2_Remove_excessive_coverage.R
# This script requires the input files
# SNPs_goodquality.txt and
# "excessive_coverage.txt".
# The output file is "Good_coverage_SNPs.txt"
```

Filter 3: remove SNPs covered in an excessive number of individuals.

SNPs that are covered in an excessive number of individuals are removed using an R script to further remove SNPs in potential repeat regions or mapping artifacts.

The number of individuals for which there is coverage for a given SNP is indicated in the last column of the .mafs file (*see* above). The cutoff is defined by the user after examining the upper tail of the distribution of the plot of the number of SNPs

against the number of individuals. SNPs in the upper tail of the distribution are discarded. Here as an example, a cutoff of ten individuals was defined.

```
# Source the "excessive_individuals" script
# (Appendix 7) file using "Rscript" or execute
# individual lines from the script in the R
# environment.
Rscript Script5_Filter3_excessive_individuals.R
# Requires input file from Script4
# "Good_coverage_SNPs.txt"
# The output file is "SNP_10ind.txt"
```

Filter 4: remove SNPs with a MAF (Minor Allele Frequency) <0.05.

The R script to discard the SNPs that have a MAF less than 0.05 would be the same as for the number of individuals. The column that is used for the filtering is "unknownEM."

```
# Source the "Remove_rare_SNPs" script
# (Appendix 8) file using "Rscript" or execute
# individual lines from the script in the R
# environment.
Rscript Script6_Filter4_Remove_rare_SNPs.R
# Requires input file from Script5
# "SNP_10ind.txt"
# The output file is "SNPs_MAF_good.txt"
```

Filter 5 (scripts 7 and 8): Select SNPs with specified length flanking sequences for primer/probe design. This step only selects SNPs that have at least the specified length of sequence on either side of the SNP, without N's.

```
# Scripts 7 and 8 (Appendix 9 and 10) must both
# be in the target directory along with the
# output of Script6 ("SNPs_MAF_good.txt") and
# the reference FASTA file.
./Script7_SNP_call_filter_GATK.sh
SNPs_MAF_good.txt ref.fa 300 output
# '300' specifies the sequence length,
# specifying 150bp on either side of the target
# SNP.
# Substitute the name of the reference sequence
# FASTA file for "ref.fa"
# Substitute a descriptive name for "output".
# This will be the base name for 2 output
# files, output.geno.fasta and output.geno.txt.
# The latter file contains columns for SNP
# locus ("chromo" from GATK output), position,
# major allele, minor allele, unknownEM (minor
# allele frequency), pu.EM (probability of
# SNP), nInd (number of individuals with data
# at SNP site).
```

Filter 6: optional validation step.

This filtering step is optional in case one wants to compare the discovered SNPs with other datasets (for example a .mafs file obtained

using a high coverage genome or coming from another geographical area). The SNPs that are shared between the datasets are identified.

```
# Source the "compare_SNP_datasets" script
# (Appendix 11) file using "Rscript" or execute
# individual lines from the script in the R
# environment.
Script9_Filter6_compare_SNP_datasets.R
# requires input files from filter Script8
# (e.g., output.geno.txt) and another .mafs
# file. These must be specified in the script
# text for importing to "data1" and "data2".
# The script assumes that the files do not have
# headers, and adds them. If you compare to a
# .mafs file that already has a header row,
# then the script must be altered so that the
# new headers are not added.
```

Filter 7: optional check for excess heterozygosity.

If SNP discovery has been conducted on a set of ≥ 10 samples (putatively from a single population), a further filter can be used to exclude SNPs that exhibit excess heterozygosity (possibly indicating presence of duplicated genes or repeats), based on significant deviation from expectations of Hardy-Weinberg equilibrium (HWE) proportions and a negative F value. This should be done only if the samples have sufficient coverage (typically $\geq 10\times$) to allow robust detection of heterozygote excess.

```
# Run genotype likelihood caller in ANGSD,
# using list of bam files in bamlist.txt.
angsd -GL 1 -out hwe_genolike_data -nThreads 5
-doGlf 2 -doMajorMinor 1 -SNP_pval 1e-6 -doMaf
2 -HWE_pval <0.05 -bam bamlist.txt
# output columns: Chromo, Position, Major,
# Minor, hweFreq, Freq, F, LRT, p-value.
```

This process is the same as inferring SNP likelihoods in Subheading 3.7 above (may take several days), except that the HWE filter is inserted to identify all SNP loci that deviate from HWE expectations at the given p-value. The output file.hwe.gz contains the loci that deviate significantly from HWE. The output can be sorted and all positive F values removed, then used to remove SNPs with significant negative F values (excess heterozygosity) from the final dataset.

```
# Decompress the ANGSD output file
gunzip hwe_genolike_data.mafs.gz

# Remove "#" from the header of the SNP list
# from Filter 4 (above).
sed -i 's/#chromo/chromo/' SNPs_MAF_good.txt
# -i indicates to edit in place (same file)

# Remove excess heterozygosity SNPs from the
# good SNPs list.
# input files: SNPs_MAF_good.txt,
```



```
# SNPs_hweExcessHet.txt.
Rscript Script11_Filter7_Remove_HWEexcessHet.R
#output = good_hwe_SNPs.txt

# Re-add "#" to the header prior to running
# Filter 5 script to extract the SNP list and
# fasta file with flanking regions.
sed -i 's/chromo/#chromo/' good_hwe_SNPs.txt
```

Filter 8: optional filter based on BLAST comparison. The output of a BLAST search can identify sequences that match closely to non-target species (e.g., contaminants), repeat regions, mitochondrial or sex chromosomes, and gene families. The output may be opened in a spreadsheet format and filtered or sorted based on the user's search words.

```
#BLAST+ search of NCBI nucleotide (nt)
# database.
blastn -db nt -query SNPs_gen0.fasta -out
SNPs_gen0_blast.out -max_target_seqs 1 -outfmt
"6 qseqid sseqid evalue length pident
salltitles" -remote
```

3.9 Applications

This approach could be applied to any species for which a reference genome or a reference of a closely related species exists to map the reads. The main advantage of this method is that it results in the discovery of a large number of SNPs. As this approach does not allow simultaneous SNP discovery and genotyping, the identified SNPs could then be target-sequenced using custom-produced SNP arrays or target enrichment capture arrays for many applications in population genomics. These SNPs would be particularly useful for applications that need a large number of SNPs such as inferences of demographic history based on the site frequency spectrum [43] or inferences of selective sweeps [44]. Samples used for SNP discovery should ideally span the geographical range of populations that will be target-sequenced to avoid ascertainment bias [10].

SNP discovery in the absence of population data can result in some false positives, where identified SNP loci are either the result of sequencing errors, or, more commonly, assembly errors that result in duplicated loci or repeat regions being combined into single assemblies. The effect of including false positives that are in fact monomorphic (false positives due to sequencing error) is relatively minor, resulting typically in a small portion of loci being uninformative. The effect of including false positives that fall in repeated regions can be much more significant, as highly repeated loci can heavily bias the read distribution in genotyping methods that rely on capture enrichment or amplification of groups of loci. Figure 3 shows the distribution of reads from a set of 384 SNP loci genotyped from multiplexed amplicons (lifetechnologies.com/ampliseqcustom) of 95 DNA samples from blue whales. The first set of 384 loci was obtained from a draft blue whale genome assembly that had not been masked for repeats. Seven of the SNPs

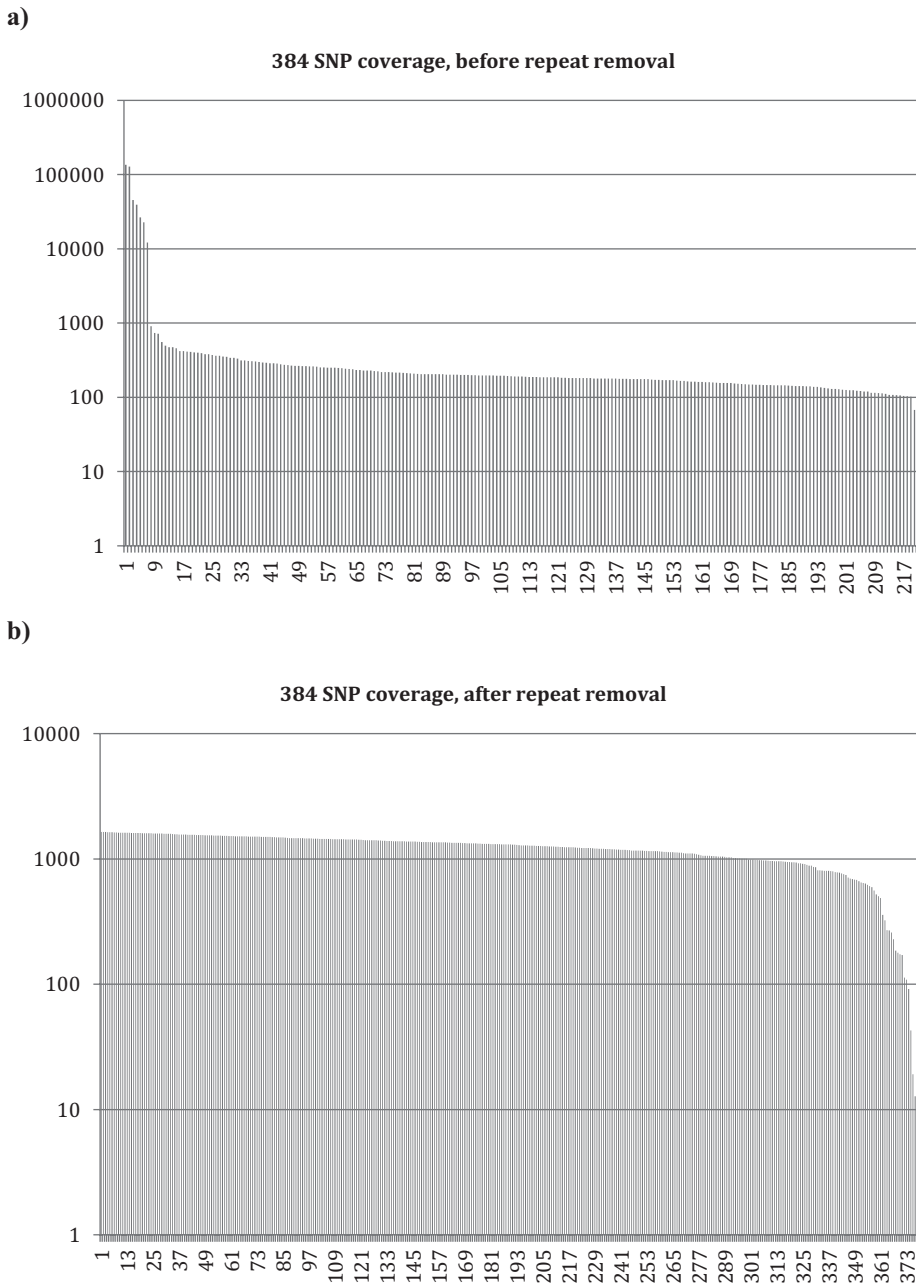


Fig. 3 Distribution of total reads assembled to 384 SNP loci **(a)** before removal of seven repeat loci, and **(b)** after removal and replacement putative repeat loci. Loci falling below the threshold coverage for SNP genotyping of any samples were excluded. The Y axis is logarithmic. Plate 1 was sequenced on an Ion Torrent (458,208 aligned reads), and plate 2 was sequenced on an Ion Proton (6,198,109 aligned reads), so the Plate 2 per-locus read depth was scaled by the ratio of plate 2 to plate 1 total reads for comparison purposes

represented 90% of the read coverage, resulting in too low coverage of the remaining loci for genotyping (*see* Fig. 3a). The SNP loci were screened for repeats using RepeatMasker as described above and replaced with SNP loci that did not show evidence of repeats. Genotyping of the second pool of 384 SNPs resulted in an almost 6× increase in average coverage of non-repeated loci and an even distribution of coverage across most loci (*see* Fig. 3b). We also recommend using BLAST to further screen for loci that either map to multiple loci or map to non-target species, e.g., bacterial or parasite DNA that could contaminate the sequence assembly.

4 Notes

1. If there are multiple sample files, running them sequentially through the same process can be facilitated by using shell scripts (e.g., `Script1_SNP_call_filter.sh`, Appendix 1). A shell script is simply a text document containing the commands that would be entered sequentially. For running the same commands on a set of sample files, the same command can be entered for each sample, with only the filenames (input and output) changed. The script is executed by typing “./” before the script name in the directory where your script and input files reside. Permissions may need to be set appropriately to allow execution of shell scripts (“`chmod`” in Linux commands, *see* Table 2).
2. Within the shell script text file, the parameters for minimum and maximum depth of coverage are set by the user, as determined for the average depth of coverage of the BAM alignment file. The script also calls the python script, which must be in the same directory, and the path to the python program may also have to be specified.
3. The filenames provided in these steps are used later in R-scripts, so changing the names will require changes in the R-scripts as well. This process is slow, and may take days to complete. It does not typically use >1 thread, though it may increase thread use as it adds sample files. A newly published wrapper for the program ANGSD [45] may make some of these steps easier or improve visualization, but has not yet been tested by us.
4. If the mapping quality filter is applied and results in an empty output file (i.e., SNPs with mapping quality <30 have all been previously filtered out), then the subsequent filter will produce an error. This can be circumvented by copying the first line of the “`excessive_coverage.txt`” file and pasting it into the empty “`poor_mapping_quality.txt`” file prior to running `Script4a`.

Table 2
Useful Linux commands

wget	wget stands for “web get”, used for non-interactive downloading of files using HTTP, HTTPS and FTP protocols.
nohup	Creates “nohup” file and relegates screen output to that file. When used with “&” at the end of the command line, the process is run in the background with no screen output. This process is incompatible with some programs and functions, so it is best to run commands from within a shell script, and use nohup/& when executing the shell script.
top	Typing “top” shows the current processes on the system, including how the percent CPU (i.e., 100% = 1 CPU) and percent of the total memory being used by each process. Exit with the keyboard combination “control” and “c”
df -h	Shows current system status, including size, amount used, amount available, % used
control-c	For a process that is active (i.e., not run in background mode), it can be cancelled with the keyboard combination “control” and “c”
kill	When a process is running in the background, it can be “killed” by using the “top” command to find the process ID (PID), then typing “kill” and the PID
Rscript	Run an R script (follow by specifying the path/name of the script)
Chmod 755	Change permissions on a file to allow everyone to read and execute the file, and the file owner is allowed to write to the file. Follow with file name
grep	Useful for extracting text. For subsetting a fasta file from a list of loci: <pre>grep -Fwf SNP_list.txt -A1 file.fasta grep -v '^--\$' >subset_file.fasta</pre> # file.fasta contains a list of sequence name texts that will be found and selected from the fasta file, along with the following DNA sequence line for each locus

This will simply remove one of the excessive coverage SNPs and allow the script to proceed.

- Calculating mean depth of coverage from the file data.depth-Global: Open the file in Excel, insert the depth of coverage from 0 to 100 in row 1 (above the depth data). Use a function to calculate the mean depth: $=\text{SUMPRODUCT}(B1:CW1 * B2:CW2) / \text{SUM}(B2:CW2)$. B1:CW1 contain the different depth of coverage from 0 to 100 and B2:CX2 the number of sites for each depth of coverage.

Acknowledgments

We are grateful to Lisa Komoroske for helpful comments on the manuscript. Blue whale DNA sequencing was generously provided by Tim Harkins and Clarence Lee, Life Technologies, Inc., and by Gerald Pao, Salk Institute for Biological Studies.

Appendix 1

Script1_SNP_call_filter.sh

Appendix 2

Script2_generate_genotype_blocks.py

Appendix 3

Script3a_Filter1_mapping_quality.R

Appendix 4

Script3b_Filter2_excessive_coverage.R

Appendix 5

Script4a_Filter1_Remove_poor_mapping_quality.R

Appendix 6

Script4b_Filter2_Remove_excessive_coverage.R

Appendix 7

Script5_Filter3_excessive_individuals.R

Appendix 8

Script6_Filter4_Remove_rare_SNPs.R

Appendix 9

Script7_SNP_call_filter_GATK.sh

Appendix 10

Script8_generate_genotype_blocks_GATK.py

Appendix 11

Script9_Filter6_compare_SNP_datasets.R

Appendix 12

Script10_plotQC.R

Appendix 13

Script11_Filter7_Remove_HWEexcessHet.R

References

- Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17:333–351. <https://doi.org/10.1038/nrg.2016.49>
- Narum SR, Campbell NR, Meyer KA, Miller MR, Hardy RW (2013) Thermal adaptation and acclimation of ectotherms from differing aquatic climates. *Mol Ecol* 22:3090–3097. <https://doi.org/10.1111/mec.12240>
- Seeb JE, Carvalho G, Hauser L, Naish K, Roberts S, Seeb LW (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Mol Ecol Resour* 11(Suppl 1):1–8. <https://doi.org/10.1111/j.1755-0998.2010.02979.x>
- Morin PA et al (2015) Geographic and temporal dynamics of a global radiation and diversification in the killer whale. *Mol Ecol* 24:3964–3979. <https://doi.org/10.1111/mec.13284>
- Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A (2012) Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol Biol Evol* 29:1917–1932. <https://doi.org/10.1093/molbev/mss086>
- Richards PM, Liu MM, Lowe N, Davey JW, Blaxter ML, Davison A (2013) RAD-Seq derived markers flank the shell colour and banding loci of the *Cepaea nemoralis* supergene. *Mol Ecol* 22:3077–3089. <https://doi.org/10.1111/mec.12262>
- Takahashi T, Sota T, Hori M (2013) Genetic basis of male colour dimorphism in a Lake Tanganyika cichlid fish. *Mol Ecol* 22:3049–3060. <https://doi.org/10.1111/mec.12120>
- Campbell NR, Harmon SA, Narum SR (2015) Genotyping-in-thousands by sequencing (GT-seq): a cost effective SNP genotyping method based on custom amplicon sequencing. *Mol Ecol Resour* 15:855–867. <https://doi.org/10.1111/1755-0998.12357>
- Aitken N, Smith S, Schwarz C, Morin PA (2004) Single nucleotide polymorphism (SNP) discovery in mammals: a targeted-gene approach. *Mol Ecol* 13:1423–1431
- Morin PA, Luikart G, Wayne RK, SNP Workshop Grp (2004) SNPs in ecology, evolution and conservation. *Trends Ecol Evol* 19:208–216. <https://doi.org/10.1016/j.tree.2004.01.009>
- Hancock-Hanser B, Frey A, Leslie M, Dutton PH, Archer EI, Morin PA (2013) Targeted multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Mol Ecol Resour* 13:254–268. <https://doi.org/10.1111/1755-0998.12059>
- Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst Biol* 61:717–726. <https://doi.org/10.1093/sysbio/sys004>
- Lemmon AR, Emme SA, Lemmon EM (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst Biol* 61:727–744. <https://doi.org/10.1093/sysbio/sys049>
- Eck SH, Benet-Pages A, Flisikowski K, Meitinger T, Fries R, Strom TM (2009) Whole genome sequencing of a single *Bos taurus* animal for single nucleotide polymorphism

- discovery. *Genome Biol* 10:R82. <https://doi.org/10.1186/gb-2009-10-8-r82>
15. Pavy N, Gagnon F, Deschenes A, Boyle B, Beaulieu J, Bousquet J (2016) Development of highly reliable in silico SNP resource and genotyping assay from exome capture and sequencing: an example from black spruce (*Picea mariana*). *Mol Ecol Resour* 16:588–598. <https://doi.org/10.1111/1755-0998.12468>
 16. Aslam ML et al (2012) Whole genome SNP discovery and analysis of genetic diversity in Turkey (*Meleagris gallopavo*). *BMC Genomics* 13:391. <https://doi.org/10.1186/1471-2164-13-391>
 17. Baird NA et al (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3:e3376. <https://doi.org/10.1371/journal.pone.0003376>
 18. Foote AD, Morin PA (2016) Genome-wide SNP data suggests complex ancestry of sympatric North Pacific killer whale ecotypes. *Heredity*. <https://doi.org/10.1038/hdy.2016.54>
 19. Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Mol Ecol* 22:2841–2847. <https://doi.org/10.1111/mec.12350>
 20. Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet* 17:81–92. <https://doi.org/10.1038/nrg.2015.28>
 21. Koepfli KP, Paten B, Genome KCS, O'Brien SJ (2015) The genome 10K project: a way forward. *Annu Rev Anim Biosci* 3:57–111. <https://doi.org/10.1146/annurev-animal-090414-014900>
 22. i5K Consortium (2013) The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered* 104:595–600. <https://doi.org/10.1093/jhered/est050>
 23. Schubert M, Lindgreen S, Orlando L (2016) AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes* 9:88. <https://doi.org/10.1186/s13104-016-1900-2>
 24. Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* 15:356. <https://doi.org/10.1186/s12859-014-0356-4>
 25. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27:2987–2993. <https://doi.org/10.1093/bioinformatics/btr509>
 26. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
 27. DePristo MA et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491–498. <https://doi.org/10.1038/ng.806>
 28. McKenna A et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303. <https://doi.org/10.1101/gr.107524.110>
 29. R Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
 30. Li H et al (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
 31. Card DC et al (2014) Two low coverage bird genomes and a comparison of reference-guided versus de novo genome assemblies. *PLoS One* 9:e106649. <https://doi.org/10.1371/journal.pone.0106649>
 32. Luo R et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18. <https://doi.org/10.1186/2047-217X-1-18>
 33. Simpson JT, Durbin R (2012) Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* 22:549–556. <https://doi.org/10.1101/gr.126953.111>
 34. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997v1 [q-bioGN]
 35. Lounsbury ZT, Brown SK, Collins PW, Henry RW, Newsome SD, Sacks BN (2015) Next-generation sequencing workflow for assembly of nonmodel mitogenomes exemplified with North Pacific albatrosses (*Phoebastria* spp.) *Mol Ecol Resour* 15:893–902. <https://doi.org/10.1111/1755-0998.12365>
 36. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 12:443–451. <https://doi.org/10.1038/nrg2986>
 37. Cammen KM, Andrews KR, Carroll EL, Foote AD, Humble E, Khudyakov JI, Louis M, McGowen MR, Olsen MT, Van Cise AM (2016) Genomic methods take the plunge: recent advances in high-throughput sequencing of marine mammals. *J Hered* 107(6):481–495

38. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
39. Kim SY et al (2011) Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics* 12:231. <https://doi.org/10.1186/1471-2105-12-231>
40. Skotte L, Korneliussen TS, Albrechtsen A (2012) Association testing for next-generation sequencing data using score statistics. *Genet Epidemiol* 36:430–437. <https://doi.org/10.1002/gepi.21636>
41. Nielsen R (2004) Population genetic analysis of ascertained SNP data. *Hum Genomics* 1:218–224
42. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* 15:1496–1502. <https://doi.org/10.1101/gr.4107905>
43. Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *PLoS Genet* 9:e1003905. <https://doi.org/10.1371/journal.pgen.1003905>
44. Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. *Genome Res* 20:393–402. <https://doi.org/10.1101/gr.100545.109>
45. Durvasula A, Hoffman PJ, Kent TV, Liu C, Kono TJ, Morrell PL, Ross-Ibarra J (2016) ANGSD-wrapper: utilities for analyzing next generation sequencing data. *Mol Ecol Resour.* <https://doi.org/10.1111/1755-0998.12578>

CleanTag Adapters Improve Small RNA Next-Generation Sequencing Library Preparation by Reducing Adapter Dimers

Sabrina Shore, Jordana M. Henderson, and Anton P. McCaffrey

Abstract

Next-generation small RNA sequencing is a valuable tool which is increasing our knowledge regarding small noncoding RNAs and their function in regulating genetic information. Library preparation protocols for small RNA have thus far been restricted due to higher RNA input requirements (>10 ng), long workflows, and tedious manual gel purifications. Small RNA library preparation methods focus largely on the prevention or depletion of a side product known as adapter dimer that tends to dominate the reaction. Adapter dimer is the ligation of two adapters to one another without an intervening library RNA insert or any useful sequencing information. The amplification of this side reaction is favored over the amplification of tagged library since it is shorter. The small size discrepancy between these two species makes separation and purification of the tagged library very difficult. Adapter dimer hinders the use of low input samples and the ability to automate the workflow so we introduce an improved library preparation protocol which uses chemically modified adapters (CleanTag) to significantly reduce the adapter dimer. CleanTag small RNA library preparation workflow decreases adapter dimer to allow for ultra-low input samples (down to approx. 10 pg total RNA), elimination of the gel purification step, and automation. We demonstrate how to carry out this streamlined protocol to improve NGS data quality and allow for the use of sample types with limited RNA material.

Key words Small RNA, Next-generation sequencing, NGS, Library preparation, Gel-free, MicroRNA, Adapter dimer, piwiRNA, Non coding RNA, tRNA

1 Introduction

Next-generation small RNA-Sequencing (sRNA-Seq) is becoming increasingly important due to its significance in the diagnostics field. sRNA-Seq has allowed us to dig deeper into the mechanisms of gene regulation by revealing many new types of noncoding small RNA, some of which have been identified and used as biomarkers for diagnostic assays [1]. Small RNA (sRNA) are noncoding RNA less than 500 nt in size and consist of microRNA (miRNA), piwiRNA (piRNA), transfer RNA (tRNA), Y RNA, and several other RNA species [2]. MicroRNA, a well-studied class of small

RNA, are 18–23 nt and play a major role in gene regulation [3, 4]. Other types of sRNA are comprised of various sizes and as more NGS data is gathered, their regulatory roles, and functions are being further classified.

Traditional library preparation protocols for sRNA are plagued with a specificity issue which is caused by the favored formation of a well known side product called adapter dimer (*see* Fig. 1, lane a). Adapter dimer is an “empty” product where the two adapters have ligated to themselves without an RNA insert in the middle which thus produces no valuable sequencing information. A difference of 20–30 nt between a tagged small RNA library and an adapter dimer makes separation of these products very difficult [3]. Therefore, protocols for small RNA library preparation include methods to reduce adapter dimer formation or improve size selection and purification of the desired target. The main solution to this problem thus far has been to perform a gel extraction of the desired tagged library. Gel purifications however result in tremendous variability, loss of desired material, and inhibit the automation of this technique. Furthermore, gel purification doesn’t completely remove all traces of adapter dimer resulting in contamination of the downstream sequencing reads. The adapter dimer challenge becomes more prominent when limiting RNA material is available. At lower RNA inputs the adapter dimer preferentially amplifies and dominates the PCR reaction leaving very little amplified library. For this reason, many commercially available kits recommend 100 ng and more recently 10 ng as the lowest RNA input to be used. This higher input requirement excludes the use of many biological type samples which are used for diagnostic testing since their RNA content is inherently low [5, 6].

Small RNA library preparation is carried out in six major steps (*see* Fig. 2): ligation of 3’ adapter to RNA [7], ligation of 5’ adapter onto half tagged library, cDNA synthesis of full tagged library, amplification and barcoding of final library, crude library analysis and quantification, and library purification. This protocol is compatible with Illumina sequencing technology.

In this chapter, we provide a detailed protocol for improved small RNA library preparation (CleanTag small RNA library preparation workflow) (*see* **Note 1**) which uses chemically modified adapters to suppress adapter dimer formation (*see* Fig. 1, lane b) and allows for (1) lower RNA inputs, (2) gel-free purification, and (3) potential for automation and higher throughput. Chemical modifications were placed on the 5’ and 3’ Illumina compatible adapters near the ligation junctions to prevent ligation of the adapters to one another but allow efficient ligation of the adapters to the library. Furthermore, the modifications help to suppress read through by the reverse transcription enzyme if any residual adapter dimer is formed. Reagent and workflow optimizations

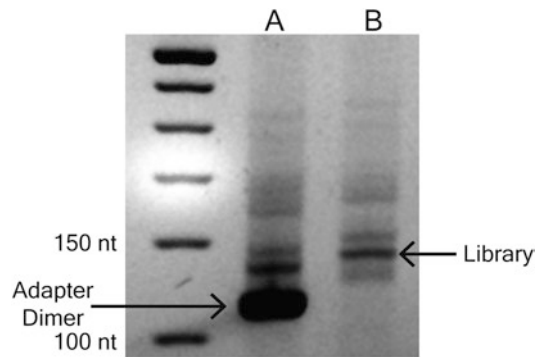


Fig. 1 Agarose gel results of a small RNA library preparation. (a) Traditional small RNA library preparation. (b) CleanTag small RNA library preparation

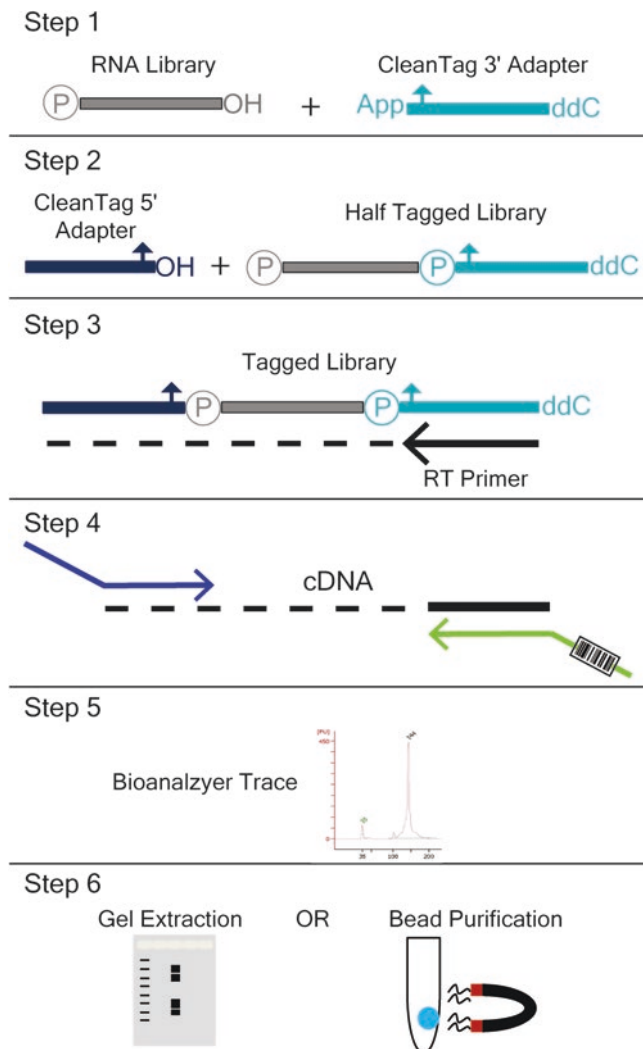


Fig. 2 Schematic of the CleanTag small RNA library preparation workflow. (Step 1) 3' adapter ligation. (Step 2) 5' adapter ligation. (Step 3) cDNA synthesis. (Step 4) PCR. (Step 5) Analysis. (Step 6) Purification

were completed to ensure optimal library yield with minimal adapter dimer formation. Libraries prepared with the CleanTag modified adapters using 1–1000 ng of total RNA and gel purified produce less than 1% adapter dimer reads while maintaining the same levels of miRNA reads at lower inputs. In addition, the CleanTag adapters have been used successfully to sequence single-cell levels of total RNA (~10 pg), a significant improvement over current protocols [8]. Even at these ultra-low levels of RNA input the adapter dimer reads are less than 3% of total reads. Collaborators have tested this kit with challenging sample types and have produced quality sequencing data from the following RNA inputs: plasma, serum, exosomes, urine, FACS sorted cells, FFPE, and immunoprecipitation assays [8, 9].

One further benefit to reducing adapter dimer formation is to eliminate the gel purification step and use a bead-based purification method to size select your library of interest. When doing bead-based purification, adapter dimer levels remain lower than 10% for inputs down to 1 ng of total RNA, a vast improvement over other protocols. Bead-based purification enables automation of this small RNA library preparation workflow and increases throughput. The CleanTag small RNA library preparation protocol has advanced sRNA-Seq technology to allow improved sensitivity, better specificity, and more streamlined workflows, even for challenging biological samples.

2 Materials

1. Nuclease-free water.
2. 200 μ L thin-walled PCR tubes.
3. 1.7 mL eppendorf tubes, pipet tips.
4. Thermal cycler with heated lid.
5. BioAnalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA) (optional, but recommended).
6. Real Time Thermal Cycler (optional).
7. Qubit fluorometer (ThermoFisher Scientific, Waltham, MA, USA) (optional).
8. Gel electrophoresis apparatus.
9. Magnetic bead separator.
10. PAGE purified oligonucleotides (Tables 1 and 2). (abbrev: ddC = 2',3'-dideoxycytidine, rApp = adenylyate, MP = methylphosphonate, 2'OMe = 2'-O-methyl, r = RNA).

2.1 3' Adapter**Ligation**

1. 2× Buffer 1: 100 mM Tris(hydroxymethyl)aminomethane–HCl (Tris–HCl) pH 7.5, 20 mM MgCl₂, 2 mM dithiothreitol (DTT), 37.5% polyethylene glycol (PEG) 8000, pH 7.6 at 25 °C.
2. CleanTag 3' Adapter (*see* Table 1), (5 μM).
3. RNase Inhibitor: murine RNase Inhibitor, 40 U/μL.
4. Enzyme 1: T4 RNA Ligase 2 truncated, KQ double mutant, 200 U/μL.
5. RNA template: User supplied. Use human brain total RNA as a positive control.

2.2 5' Adapter**Ligation**

1. 10× Buffer 2: 500 mM Tris–HCl pH 7.5, 100 mM MgCl₂, 10 mM DTT, 20 mM ATP, pH 7.6 at 25 °C.
2. CleanTag 5' Adapter (*see* Table 1), (20 μM).
3. RNase Inhibitor: murine RNase Inhibitor, 40 U/μL.
4. Enzyme 2: T4 RNA Ligase 1, 10 U/μL.

2.3 cDNA Synthesis

1. 5× Protoscript II reverse transcription buffer: 250 mM Tris–HCl, 375 mM KCl, 15 mM MgCl₂, pH 8.3 at 25 °C.
2. Dithiothreitol (DTT), (100 mM).
3. RNase Inhibitor: murine RNase Inhibitor, 40 U/μL.
4. Protoscript II Reverse Transcriptase (RT), 200 U/μL.
5. Reverse Transcription Primer, 5 μM (*see* Table 1).
6. Deoxyribonucleotide triphosphates (dNTPs), 10 mM each.

2.4 PCR

1. 2× High Fidelity PCR Master Mix: Q5 High Fidelity PCR Master Mix (New England Biolabs, Ipswich, MA, USA).
2. Forward primer, 20 μM (*see* Table 1).
3. Reverse Index Primer, 20 μM (*see* Tables 1 and 2).

2.5 Analysis (See Note 2)

1. High Sensitivity DNA chip for Agilent Bioanalyzer (Agilent Technologies).
2. 4% Agarose EX Gels (ThermoFisher Scientific).
3. KAPA Quantitative library qPCR kit for Illumina Sequencing (KAPA Biosystems, Wilmington, MA, USA).
4. Qubit dsDNA HS assay (ThermoFisher Scientific).

2.6 Purification**2.6.1 Bead-Based Purification**

1. Agencourt AMPure XP beads (Beckman Coulter, Carlsbad, CA, USA).
2. 70% Ethanol.
3. Magnetic Rack.

Table 1
Clean Tag small RNA library preparation oligonucleotide sequences

Name	Sequence 5' to 3'
Clean Tag 3' Adapter	(rApp)T(MP)GGAATTCTCGGGTGCCAAGG (ddC)
Clean Tag 5' Adapter	r[GUUCAGAGUUCUACAGUCCGACGAU (C) 2'OMe]
RT primer	GCCTTGGCACCCGAGAAATCCCA
Forward primer	AATGATACGGCGACCCACCGAGATCTACACGTTTCAGAGTTCTACAGTCCGA
^a Reverse primer (Index Primer)	CAAGCAGAAAGACGGGCATACGAGAT <u>CGTIGAT</u> GTGACTGGAGTTCCTTGGCACCCGAGAAATCCCA

^aUnderlined sequence represents unique barcode that is interchangeable with sequences listed below

Table 2
Illumina barcoding sequences used in index primers (#1–24)

Index 1	CGTGAT	Index 7	GATCTG	Index 13	TTGACT	Index 19	TTTCAC
Index 2	ACATCG	Index 8	TCAAGT	Index 14	GGAACT	Index 20	GGCCAC
Index 3	GCCTAA	Index 9	CTGATC	Index 15	TGACAT	Index 21	CGAAAC
Index 4	TGGTCA	Index 10	AAGCTA	Index 16	GGACGG	Index 22	CGTACG
Index 5	CACTGT	Index 11	GTAGCC	Index 17	CTCTAC	Index 23	CCACTC
Index 6	ATTGGC	Index 12	TACAAG	Index 18	GCGGAC	Index 24	GCTACC

2.6.2 Gel Purification

1. Zymoclean Gel DNA Recovery Kit (Zymo Research, Irvine, CA, USA).
2. Scalpel.

3 Methods

It is important to plan out the entire sequencing experiment before you begin. Identify starting amount of RNA to adjust protocol as needed (*see Note 3*). Consider how many samples you have and choose appropriate barcodes for multiplexing purposes (*see Note 4*). Avoid introducing RNases by using RNase-free laboratory techniques. Do not vortex enzymes and always keep them on ice. After reagents have been thawed, place them immediately on ice. Keep PCR tubes with reaction components on ice at all times.

3.1 Ligation of CleanTag 3' Adapter to RNA Template

1. Thaw all the necessary reagents for Subheading 3.1.
2. Determine if modified 3' adapter requires dilution based on available RNA input as suggested in **Note 3**. If necessary dilute adapter in nuclease-free water and mix well before use. Dilute enough adapter for all samples plus 10% excess if preparing master mixes (*see Note 5*). Each sample requires 1 μ L CleanTag 3' adapter.
3. Prepare a reaction mixture by adding the following components into thin-walled PCR tubes in the following order: 1 μ L nuclease-free water, 1 μ L CleanTag 3' adapter, 1 μ L RNase Inhibitor, 1 μ L Enzyme 1, and 5 μ L Buffer 1, per reaction. Buffer 1 is extremely viscous, pipet slowly. Keep mix cold. If preparing multiple reactions refer to **Note 5** for tips on making a master mix and add 9 μ L of the master mix to individual reaction tubes. Mix the reaction by pipetting up and down and pulse spin.

4. Heat RNA template for 2 min at 70 °C in thermal cycler with heated lid to denature any secondary structure. Immediately place template on ice.
5. Add a minimum of 1 µL RNA to each individual reaction tube for a total of 10 µL (*see Note 6*). Mix by gently pipetting volume up and down then pulse spin.
6. Incubate the reaction for 1 h at 28 °C followed by 20 min at 65 °C in thermal cycler with heated lid for ligation and subsequent enzyme inactivation.
7. Place half tagged library sample on ice.

3.2 CleanTag 5' Adapter Ligation to Half Tagged Small RNA Library

1. Thaw all the necessary reagents for Subheading 3.2.
2. Determine if dilution of CleanTag 5' adapter is necessary. Refer back to **Note 3** and make any dilutions necessary in nuclease-free water. Be sure to prepare enough adapter for all samples with 10% extra to account for pipetting error. Each 5' adapter ligation reaction requires 2 µL CleanTag 5' adapter.
3. Add reagents in the following order to the thin-walled PCR tubes which contain the ligation mixture from Subheading 3.1: 4 µL nuclease-free water, 1 µL Buffer 2, 1 µL RNase Inhibitor, and 2 µL Enzyme 2 per reaction. Keep cold and do not vortex enzymes. If preparing multiple reactions a master mix with 10% excess of each reagent can be mixed and 8 µL aliquoted to individual reaction tubes (*see Note 5*). Mix the reaction by pipetting up and down and pulse spin.
4. Heat CleanTag 5' adapter for 2 min at 70 °C in thermal cycler with heated lid. Immediately place on ice.
5. Add 2 µL denatured CleanTag 5' adapter to reaction mix for a total of 20 µL. Mix by pipetting.
6. Incubate for 1 h at 28 °C followed by 20 min at 65 °C.
7. Place full tagged library sample on ice.

3.3 cDNA Synthesis of Full Tagged Library

1. Thaw all necessary reagents for Subheading 3.3.
2. Add 2 µL RT Primer to full tagged library product from Subheading 3.2 for a total of 22 µL. Mix gently by pipetting up and down.
3. Place tubes in thermal cycler with heated lid and incubate for 2 min at 70 °C. Immediately place on ice.
4. Add the following to the thin-walled PCR tubes: 1.92 µL nuclease-free water, 5.76 µL 5× RT Buffer, 1.44 µL 10 mM dNTP mix, 2.88 µL DTT, 1 µL RNase Inhibitor, and 1 µL RT Enzyme per reaction for a total of 36 µL. Mix by pipetting up and down and keep mix cold. If performing multiple reactions

a master mix with 10% excess of each reagent can be prepared and 14 μL added to individual reaction tubes (*see* **Note 5**).

5. Incubate for 1 h at 50 °C in thermal cycler with heated lid.

3.4 Amplification and Barcoding of Final Library

1. Determine proper index barcode primers to use based on the number of library samples to be sequenced (*see* **Note 4**).
2. Determine appropriate number of PCR cycles to use based on RNA input as suggested in **Note 3**.
3. Thaw necessary reagents for Subheading 3.4.
4. Add the following to the thin-walled PCR tubes from Subheading 3.3: 40 μL 2 \times High Fidelity PCR Master Mix and 2 μL Forward Primer per reaction (*see* **Note 7**). Mix gently by pipetting up and down. Keep cold. If performing multiple library reactions a master mix with 10% excess of each reagent can be prepared and 42 μL of master mix can be added to individual reaction tubes.
5. Add 2 μL appropriate Index Reverse Primer to each individual tube for a total of 80 μL . Mix gently by pipetting up and down.
6. Place tubes in thermal cycler with heated lid and run the following cycling conditions: 98 °C for 30 s; x cycles of 98 °C for 10 s, 60 °C for 30 s, 72 °C for 15 s; 72 °C for 10 min, then 4 °C hold, where x is number of PCR cycles as determined by user.
7. Proceed to analysis and quantification of crude library or place samples at -20 °C until ready for use.

3.5 Crude Library Analysis and Quantification

Several options exist for crude library analysis and quantification including (1) Agilent Bioanalyzer by High Sensitivity DNA Chip (Recommended), (2) qPCR (Library Quantification Kit for Illumina platforms), or (3) gel electrophoresis. Visualization by gel electrophoresis is an alternative method if Bioanalyzer or qPCR are not available but gel images are difficult to quantify precisely and can lead to improper pooling of samples. As well, large sample volumes are often necessary for visualization on gel which could consume valuable library sample before being purified and sequenced.

3.5.1 Crude Library Analysis and Quantification by Agilent Bioanalyzer

This option is best for visualization of all library peak sizes while conserving the majority of your sample for purification and sequencing. It is important to follow the manufacturer's instructions for the proper use and maintenance of this instrument. Clean electrode pins and change syringe in priming station often.

1. Vortex and spin down PCR products before use. Dilute samples 1:10 by combining 2 μL PCR product to 18 μL nuclease-free water. Vortex well and collect liquid at bottom of tubes.

2. Prepare High Sensitivity DNA chip for analysis. Follow the manufacturer's instructions.
Load 1 μL of diluted sample into chip per the manufacturer's recommended instructions.
3. Analyze Results. MicroRNA libraries should run at approximately 140 nt. Adapter dimer, if present, will run at ~ 120 nt and a library peak at 150 nt indicates a small RNA of ~ 30 nt (this could be piwiRNA, tRNA, etc. depending on sample input) (*see* Fig. 3). Given this, sometimes adjustments may need to be performed manually on the BioAnalyzer to generate accurate peak sizes (*see* Note 8 and Fig. 4).
4. Determine best purification method for your experiment such as (a) bead-based size selection or (b) gel purification (*see* Note 9). For bead-based size selection, if crude library meets expected criteria, proceed to Subheading 3.6.1. For gel purification it is important to quantify each library prior to pooling samples together for gel extraction. Equimolar pooling for multiplexed sequencing is critical (*see* Note 10). If analyzing by Bioanalyzer, perform Smear Analysis as per the manufacturer's instructions using 2100 Expert software to quantify library peak. Be sure to quantify only peaks of interest that you plan to extract from the gel within smear settings.

3.5.2 Crude Library Analysis by qPCR

Several kits are available for quantitative analysis of Illumina compatible libraries. It is important to follow the manufacturer's instructions in order to obtain accurate results. Library quantification of individual crude samples will be necessary prior to gel purification if pooling samples together.

3.6 Library Purification

Multiple options exist for library purification, two of which are: (1) bead-based size selection or (2) gel extraction. Identify the best purification method for your experiment (*see* Note 9).

3.6.1 Bead-Based Size Selection

1. Bring AMPure XP beads to room temperature for 30 min. Vortex well to ensure a homogenized mixture before use. Transfer library samples from 0.2 mL PCR tubes into 1.7 mL microcentrifuge tubes to account for larger volumes used in this protocol.
2. Add 1 \times volume of AMPure XP beads to library. For example, use 80 μL beads with 80 μL PCR product. Vortex well to mix beads and DNA, then spin down briefly to collect liquid at the bottom of the tube.
3. Incubate the tubes at room temperature for 10 min.
4. Place the tubes on magnetic rack for 4 min to separate the beads onto the side of the tube nearest the magnetic rack. The supernatant will become clear.

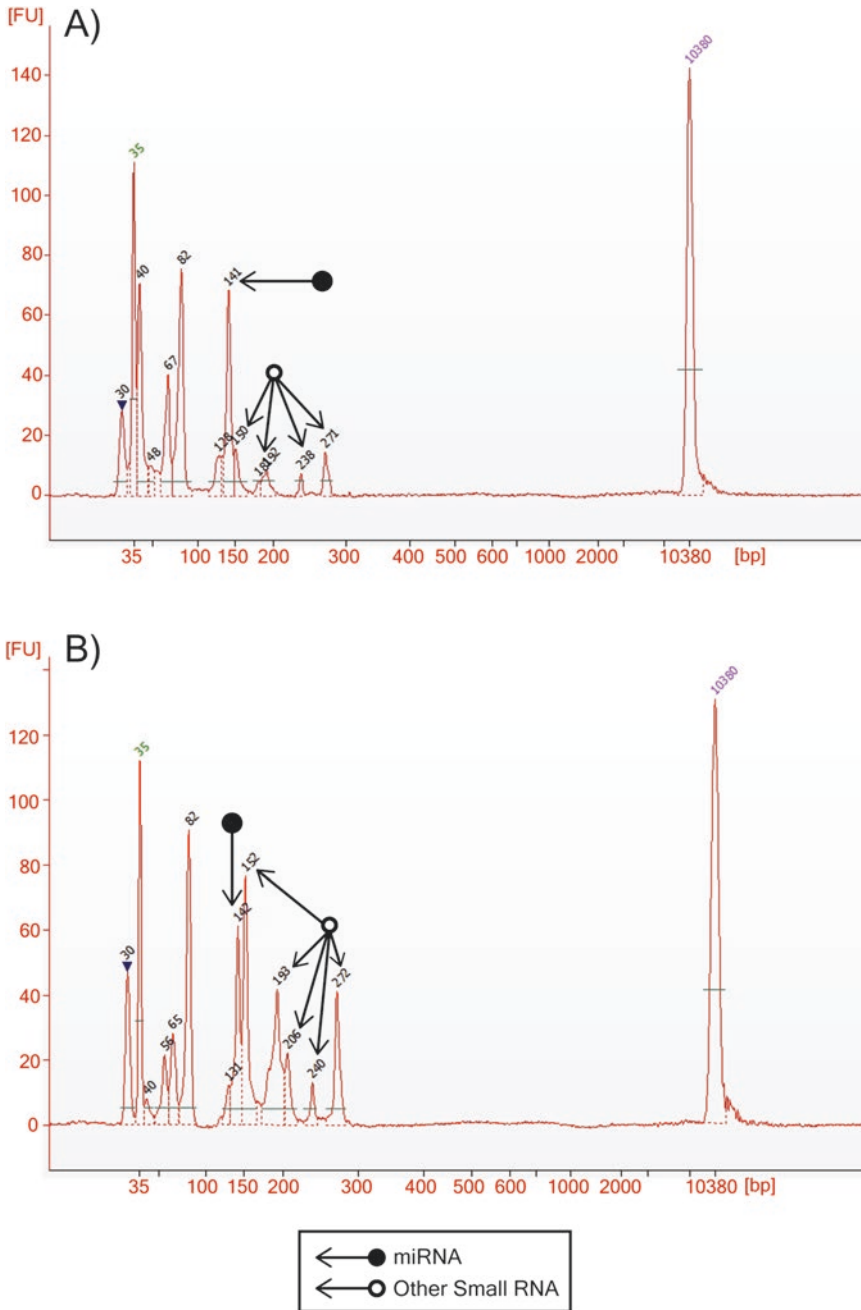


Fig. 3 Bioanalyzer traces of crude small RNA libraries from total human brain RNA at **a**) high input (1000 ng) miRNA at 141 nt, or **b**) low input (1 ng) miRNA at 142 nt

5. Keep the tubes on magnetic rack and carefully transfer the supernatant to clean new tube off rack without disrupting magnetic bead pellet on the side of the original tube wall. Discard the beads, these contain higher molecular weight species above ~200 nt.

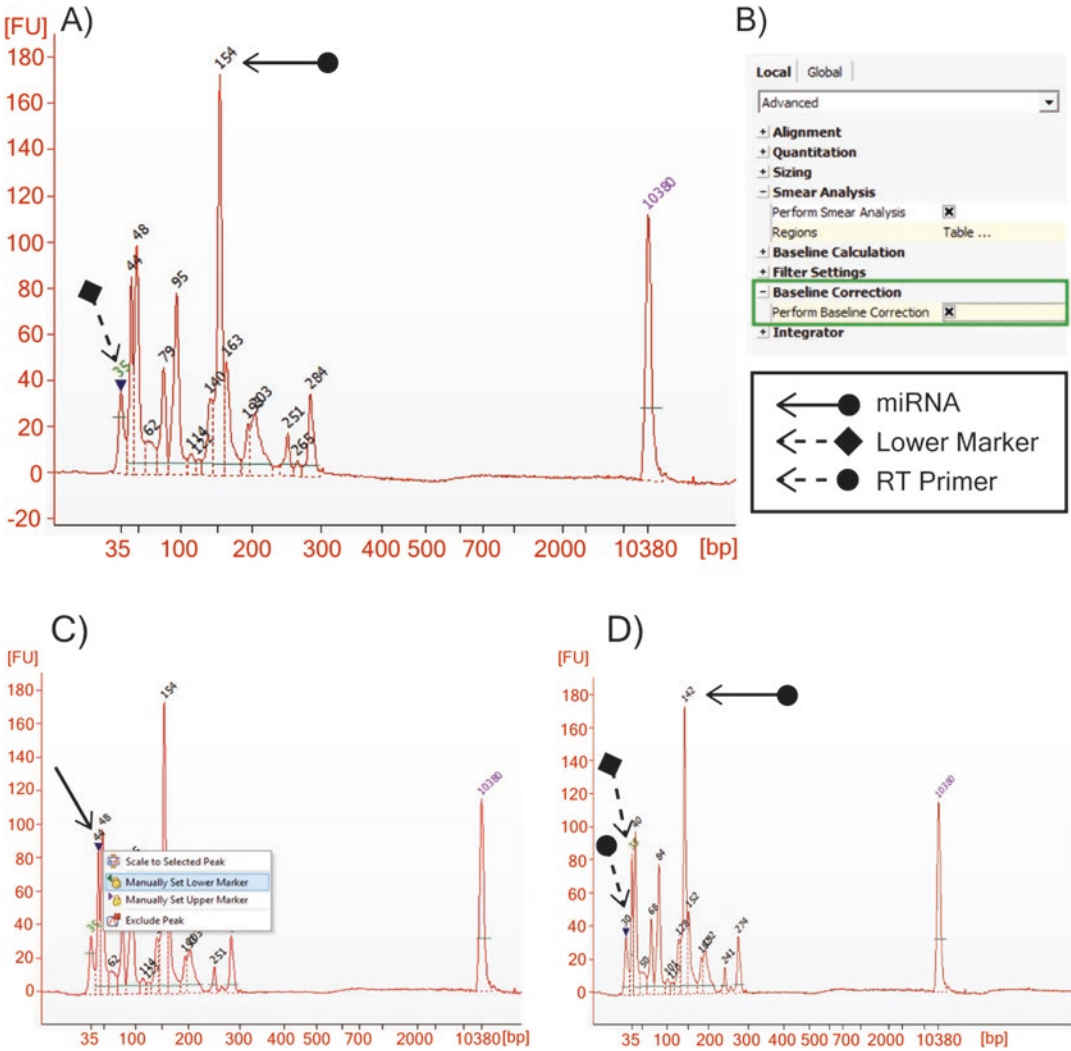


Fig. 4 Example of manual adjustments in 2100 Expert BioAnalyzer software. (a) Crude trace with miRNA library at 154 nt and lower marker selected as the first peak. (b) Baseline correction adjustment. (c) Manually adjust lower marker to the second peak. (d) Crude trace with miRNA library at true size of 142 nt and RT primer at 30 nt (first peak)

6. Add 1.8× the original PCR volume (144 μ L if started with 80 μ L) AMPure XP beads to transferred supernatant. Vortex well to mix thoroughly and spin briefly to collect liquid at the bottom of the tube.
7. Incubate off magnetic rack at room temperature for 10 min.
8. Place the tubes on magnetic rack for 4 min to separate the beads onto the side of the tube near rack. The supernatant will become clear.
9. Keep the tubes on magnetic rack and aspirate the supernatant carefully without disrupting bead pellet. Discard the supernatant, library of interest (100–200 nt) is now bound to beads.

10. Wash the beads while on magnetic rack by gently adding 500 μL 70% ethanol into the tubes. Wait 30 s and then collect wash without disrupting bead pellet and discard ethanol appropriately. Repeat for a total of two washes. After second wash remove all traces of ethanol.
11. Air-dry bead pellet while on magnetic rack for 5 min. Do not over-dry beads but ensure all ethanol has evaporated.
12. Resuspend bead pellet in 17 μL nuclease-free water and incubate for 2 min off magnetic rack at room temperature.
13. Place the tubes on magnetic rack for 2 min or until the beads and the supernatant have separated.
14. Keep the tubes on magnetic rack and collect 15 μL of supernatant without disrupting bead pellet and transfer to a clean tube. This supernatant contains the library. Store at 4 $^{\circ}\text{C}$ or -20°C long term.
15. Reanalyze purified libraries on Bioanalyzer (*see* Fig. 5a). Quantify using a Qubit or perform smear analysis between 100–300 nt on the Bioanalyzer. For multiplexed sequencing on Illumina platforms pool libraries together for an equimolar mix of all samples (*see* Note 10).

3.6.2 Gel Extraction

1. Pool individual samples equally into one mixture according to Note 10.
2. Load mixture onto a 4% EX Agarose Gel (or best available gel option). If the sample does not fit into one lane, split the sample into several lanes. Be sure to include a known marker or ladder in one of the lanes for size reference. Run gel for 30 min and then visualize band safely using proper eye protection against UV.
3. Using scalpel cut out target of interest.
4. Follow Zymoclean Gel DNA Recovery Kit manufacturer's instructions for gel purification procedure. Note: May use alternative gel purification kits if desired.
5. Quantify purified samples. We recommend visualizing sample with Bioanalyzer to ensure library of interest was properly extracted (*see* Fig. 5b). At this point, Qubit may be used to quantify samples or pools (*see* Note 11).

4 Notes

1. CleanTag Adapters and their use are covered by one or more patents or pending patent applications owned by TriLink BioTechnologies, LLC. Chemically Modified Ligase Cofactors, Donors and Acceptors (WO 2014144979 A1; US 8728725B2; 20140323354).

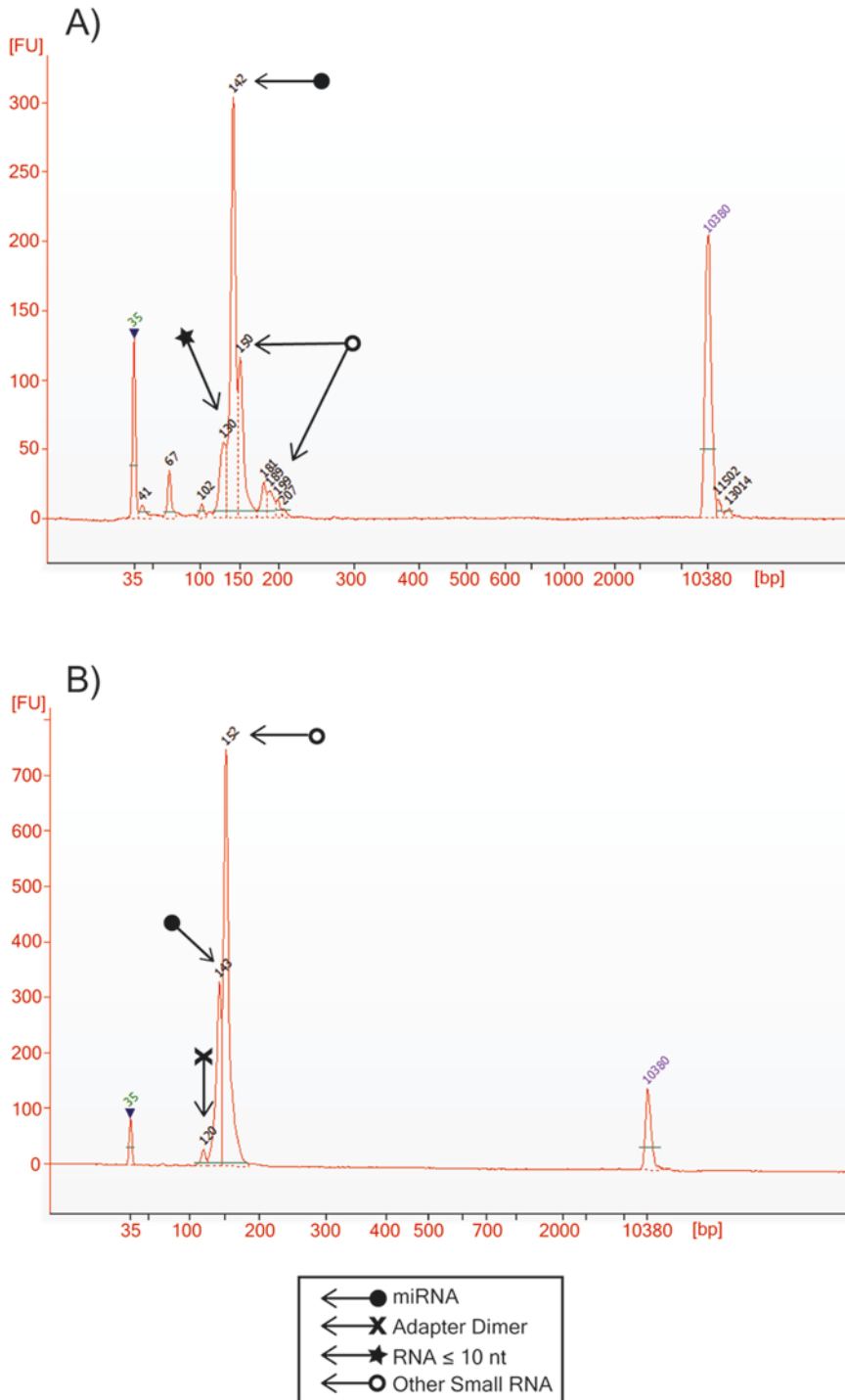


Fig. 5 Purified small RNA libraries from total human brain RNA. **(a)** Bead-based purification of one sample at 1000 ng input. **(b)** Gel purification of six ultra-low input (100–10 pg) samples which were pooled together prior to gel purification

2. Reagents and kits of comparable performance may be substituted for the following recommended items.
3. Determine starting amount of RNA input. For low inputs, adapters should be diluted before use and number of PCR cycles should be increased for best results. Values in Table 3 are recommendations based on experiments using high-quality total human brain RNA as an input. Low-quality samples may actually require further dilution of adapters and or increased PCR cycles. Samples that are enriched for miRNA or other small RNA may require some experimental optimization. Use Table 3 only as a guide.
4. Identify number of samples for full NGS experiment. When fewer libraries are pooled together certain barcodes must be used during library amplification in order to introduce proper diversity needed per lane of flow cell during sequencing. Follow the manufacturer's recommended instructions.
5. For library preparation of multiple samples we recommend making a master mix at each step with 10% excess of each component. For example, in Subheading 3.1, **step 3**, 1 μL of nuclease-free water is added to each reaction. If you are preparing ten libraries use 11 μL water in the master mix as calculated by $1 \mu\text{L per sample} \times 10 \text{ samples} \times 1.1$ (for the 10% excess) equals 11 μL . To complete the master mix in Subheading 3.1, **step 3**, add 11 μL CleanTag 3' adapter, 11 μL RNase Inhibitor, 11 μL Enzyme 1, and 55 μL Buffer 1 to the 11 μL water and mix thoroughly. When master mix is ready, only aliquot the sum volume for one reaction into tubes. For example, a single reaction in Subheading 3.1, **step 3**, has a volume of 9 μL before adding the RNA template, therefore, aliquot 9 μL of master mix into the ten different library preparation tubes.
6. For samples with low concentration, up to 10 μL RNA can be added to 3' ligation reaction. Other reagents need not be adjusted even though volumes will be different.
7. White precipitate may appear in 2 \times High Fidelity PCR Master Mix, gently pipet up and down to homogenize before use. Do not heat master mix above $-20 \text{ }^\circ\text{C}$ for prolonged periods of time.
8. High concentrations of primers and adapters are used to push reactions forward during library preparation. In a crude sample these excess primers will appear as peaks on the High Sensitivity DNA chip analysis. Of particular importance the RT Primer runs very close in size to the lower marker at 35 nt which can cause the marker to shift and peak sizes will be incorrectly assigned. Manually set lower marker as the second of three front peaks if necessary as shown in Fig. 4 by right clicking on peak for drop down menu.

Table 3
Optimization table based on total RNA input: adapter dilution and PCR cycles

Total RNA input	Adapter dilution	PCR cycles
1000 ng	1×	12×
100 ng	1:2×	15×
10 ng	1:4×	18×
1 ng	1:12×	21×
100 pg	1:14×–1:18×	24×
10 pg	1:18×–1:22×	27×

9. Depending on your samples and experimental design you may benefit most from different purification methods. Bead-based purification is automatable for higher throughput, however this method is not recommended if minimal levels of adapter dimer have formed as the size of library and dimer are too close for separation. The range of size selection between 100 to 200 nucleotides when using the bead-based method allows for the inclusion of multiple small RNA species to be sequenced and analyzed. Please note that if you'd like to analyze longer RNA fragments you can do a PCR cleanup to eliminate products less than 100 bp and keep all higher molecular weight targets. For this you would start at step 6 in Section 3.6.1 by adding 1.8× beads directly to the PCR sample and finish out the protocol. With bead-based size selection the samples can be pooled together after purification allowing for more accurate quantification analysis. Alternatively, gel purification can be used for the extraction of a specific small RNA library and samples can be pooled together and purified at the same time.
10. To multiplex samples for sequencing within a single lane on a flow cell it is important to add equimolar amounts of each sample to have an even representation of each library during data acquisition. First identify the molarity of individual libraries (usually in nM). Refer to sequencing platform requirements to next determine necessary volume and concentration of final multiplexed pool. For example, if the final pool should be a minimum of 10 nM in 25 µL volume and you have 24 indexed libraries then you will need to pool 0.416 nM each sample. Because you want each of the 24 samples to occupy equal space on the flow cell each sample should make up 4.16% of the total, or $100\% / 24 = 4.16\%$. Next, since the total concentration will be 10 nM each sample should make up 4.16%

of 10 or 0.416 nM. Then calculate the volume necessary per sample based on individual concentrations into a final volume of 25 μ L. For example, if sample 1 is at 10 nM a volume of 1.04 μ L should be pooled for a final of 10 nM in 25 μ L with 24 samples where $(0.416 \text{ nM} * 25 \mu\text{L})/10 \text{ nM}$ gives a volume of 1.04 μ L. Some samples may need to be diluted before pooling in order to pipet a manageable volume. If necessary, bring final pool up to volume using nuclease-free water after all 24 samples have been mixed together.

11. Qubit fluorometer can be used to quantify purified samples with the dsDNA high sensitivity (HS) assay. This method uses intercalating dyes so anything double stranded will contribute to the total concentration read out. For this reason we only recommend using a Qubit after libraries have been purified and unwanted products have been eliminated. Follow the manufacturer's recommended instructions.

References

1. Lopez JP et al (2015) Biomarker discovery: quantification of microRNAs and other small non-coding RNAs using next generation sequencing. *BMC Med Genet* 8:35
2. Huttenhofer A, Brosius J, Bachellerie JP (2002) RNomics: identification and function of small, non-messenger RNAs. *Curr Opin Chem Biol* 6(6):835–843
3. Head SR et al (2014) Library construction for next-generation sequencing: overviews and challenges. *Biotechniques* 56(2):61–64. 66, 68, passim
4. Vickers KC et al (2015) Mining diverse small RNA species in the deep transcriptome. *Trends Biochem Sci* 40(1):4–7
5. Etheridge A et al (2011) Extracellular microRNA: a new source of biomarkers. *Mutat Res* 717(1–2):85–90
6. Sterling CH, Veksler-Lublinsky I, Ambros V (2015) An efficient and sensitive method for preparing cDNA libraries from scarce biological samples. *Nucleic Acids Res* 43(1):e1
7. Song Y, Liu KJ, Wang TH (2014) Elimination of ligation dependent artifacts in T4 RNA ligase to achieve high efficiency and low bias microRNA capture. *PLoS One* 9(4):e94619
8. Shore S, Henderson JM, Chow FWN, Tam EWT, Quintana JF, Chhatbar K, McCaffrey AP, Zon G, Buck A, Hogrefe RI (2016) Improved small RNA library preparation for ultra-low input sample types, in NGS, SCA, SMA, and Mass Spec: research to diagnostics, San Diego, CA. <http://www.trilinkbiotech.com/work/sRNALibPrep.pdf>
9. Shore S, Henderson JM, Quintana JF, Chow FWN, Chhatbar K, McCaffrey AP, Zon G, Buck A, Hogrefe RI (2016) Improving NGS small RNA discovery in biological fluids and other low input samples, in keystone symposia—small RNA silencing: little guides, big biology (A6), Keystone, CO. <http://www.trilinkbiotech.com/work/sRNASilencing.pdf>

Chapter 11

Sampling, Extraction, and High-Throughput Sequencing Methods for Environmental Microbial and Viral Communities

Pedro J. Torres and Scott T. Kelley

Abstract

The emergence of high-throughput sequencing technologies has deepened our understanding of complex microbial communities and greatly facilitated the study of as-yet uncultured microbes and viruses. Studies of complex microbial communities require high-quality data to generate valid results. Here, we detail current methods of microbial and viral community sample acquisition, DNA extraction, sample preparation, and sequencing on Illumina high-throughput platforms. While using appropriate analytical tools is important, it must not overshadow the need for establishing a proper experimental design and obtaining sufficient numbers of samples for statistical purposes. Researchers must also take care to sample biologically relevant sites and control for potential confounding factors (e.g., contamination).

Key words 16S/18S ribosomal RNA gene sequences, DNA extraction, Microbiome, Viral purification, Viral-like particles (VLPs)

1 Introduction

The first step in the analysis of any microbial or viral community is to establish the necessary infrastructure for the collection and storage of samples. The preparation, handling, and storage of samples is a critical, and often underrated, process that can alter the outcomes of downstream DNA-based microbial community analysis. Sterile collection methods are vital, particularly when working with original sample material (e.g., soil, water, feces, tissue, etc.). Samples should be processed for DNA extraction as soon as possible, or immediately stored at ultra-low temperatures to preserve the original composition of the microbiome and its constitutive components (DNA, RNA, proteins, etc.). Shipping or transport should be completed using dry ice. Repeated thawing and freezing is particularly damaging to the integrity of microbiome samples.

1.1 *Sampling Handling*

Proper sample handling is essential for avoiding microbial growth and/or death post-collection, which could significantly alter the organismal and molecular composition and result in erroneous conclusions. Improper handling of the samples, such as exposing them for extended periods at room or hotter temperatures, promotes growth of aerobic microbes while hindering anaerobic microbial survival [1]. Other forms of improper handling, such as not using gloves or protective gear when collecting samples, using unsterilized equipment to gather samples, or repeated use of once sterile or non-sterile equipment, may result in permanent sample contamination. It goes without saying that every effort should be made to use sterile technique and eliminate potential sources of contamination during the collection phase.

1.2 *Preservation*

A common preservation technique is to immediately ultra-freeze samples post-collection (e.g., $-80\text{ }^{\circ}\text{C}$) and keep them frozen until processing. It is also possible to add a cryoprotectant for additional protection if DNA extraction will not be undertaken for a few months or longer (e.g., PBS/Glycerol 50:50 v/v). If conducting field research and freezers are unavailable, preserving in $\geq 95\%$ ethanol or buffer, depending on sample type, can be done prior to DNA extraction and analysis [2, 3]. In the case of ethanol storage, you will need to remove as much of the ethanol as possible prior to DNA isolation. There are numerous buffer kits and ethanol preservation protocols published online and all these will depend on the type of sample and time until DNA extraction will occur [2, 3].

2 *Materials*

2.1 *Sample Collection and DNA Extraction of Oral Microbiome*

1. 50 mL conical tube.
2. DNeasy PowerSoil Kit (Qiagen, Hilden, Germany).

2.1.1 *Sample Collection Oral Microbiome: Swabbing Gums, Cheeks, Tongue, Tooth Surface, or Throat*

1. Sterile cotton-tip collection swab.
2. 20 mL conical tube.

Alternate Protocol

1. Sterile water.
2. 50 mL conical tube.

2.2 *Sample Collection and DNA Extraction of Skin Microbiome*

1. Sterile nylon flocked swabs.
2. 5 mL cryovial and buffer solution.
3. Buffer options: Tris-HCl, EDTA, and 0.5% Tween 20 or 0.15 M NaCl with 0.1% Tween 20.
4. DNeasy PowerWater Kit (Qiagen, Hilden, Germany).

**2.3 Sample
Collection and DNA
Extraction of Vaginal
Microbiome**

1. Sterile swabs.
2. Cryovial.
3. DNeasy PowerSoil Kit (Qiagen, Hilden, Germany).

**2.4 Sample
Collection and DNA
Extraction of Gut
Microbiome**

1. 50 and 15 mL conical tubes.
2. Sterile spatula/plastic spoons.
3. Disposable commode specimen containers (Claflin Medical Equipment, Warwick, RI, USA).
4. Sterile spatula or plastic spoon.
5. Culture swabs (BD Biosciences, San Diego, CA, USA).
6. Single use containers, such as Solo-32 ounce poly-coated paper food container with lid.
7. Sterile forceps.
8. $\geq 95\%$ ethanol.
9. DNeasy PowerSoil Kit (Qiagen, Hilden, Germany) (*see Note 1*).

**2.5 16S rRNA
Amplicon PCR**

1. 515 forward primer (*see Notes 2 and 3*) [4].
2. 805 reverse primers (*see Notes 2 and 3*) [4].
3. PCR-grade water.
4. Thin-walled PCR tubes.
5. Platinum Hot Start PCR Master Mix (2 \times) from ThermoFisher.

**2.6 Viral Purification
and Extraction**

1. Chloroform.
2. Coral blastate.
3. Heidolph SilentCrusherTM (Heidolph North America, Elk Grove Village, IL, USA).

**2.6.1 Soil, Animal
Tissues, or Clinical
Samples**

1. Phosphate-buffered saline (PBS): 10 mM phosphate buffer, 2.7 mM KCl, and 137 mM NaCl, pH 7.4 or Saline magnesium (SM) buffer: 1 M NaCl, 10 mM MgSO₄, 50 mM Tris-HCl; adjust pH to 7.4. Filter sterilize (0.02 μ m pore size) and store at room temperature.
2. Mortar and pestle, or a homogenizer.

**2.6.2 Concentration
of Viral-like Particles**

1. Centrifuge.
2. 0.45 μ m and 0.22- μ m Whatman filter (Whatman-GE Healthcare Life Sciences, Pittsburgh, PA, USA).

**2.6.3 Density Gradients
and Ultracentrifugation**

1. Ultracentrifuge tubes.
2. Syringe.
3. 18-Gauge needle.

4. CsCl or sucrose.
5. Sterile 1.5 mL centrifuge tube.

2.6.4 Nucleic Acid Extraction

1. DNase I.
2. QIAamp MinElute Virus Spin Kit and RNeasy plus Kit (Qiagen, Hilden, Germany).
3. Nanodrop (ThermoFisher Scientific, Waltham, MA, USA) or Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA).
4. PCR primers for 16S and 18S rRNA genes for quality control (*see* **Notes 2** and **4**).

2.6.5 Random Amplification

1. Genomiphi amplification kit (GE Healthcare Life Sciences, Pittsburgh, PA, USA).
2. Qiagen DNeasy Kit (Qiagen, Hilden, Germany).

3 Methods

3.1 Sample Collection and DNA Extraction of Oral Microbiome

Sample collection of saliva is noninvasive and straightforward. To collect the saliva from the patient, have subjects spit or drool into a collection tube (50 mL conical tube) [5]. To increase yield, saliva production can be stimulated mechanically by rubbing the cheeks, just behind your back teeth.

1. Instruct the patient not to eat/drink/smoke/chew gum for at least 30 min prior to giving a sample.
2. Open 50 mL conical tube.
3. Have patient spit into collection tube until the amount of liquid saliva (not counting the foam) is at the 2 mL line.
4. Screw the lid tightly onto the collection tube and store the samples at -80°C until the samples are ready to be processed.
5. Using the MOBIO PowerSoil Isolation Kit according to the manufacturer's instruction DNA can be extracted with the following modification.
6. For saliva: add a minimum of 60 μL of saliva into bead tubes and follow the manufacturer's protocol.
7. For oral wash: Aliquot 1.5 mL of sample, centrifuge at $10,000 \times g$ and retrieve pellet for DNA isolation.
8. For swabs: break the cotton tips of the frozen swabs directly into tubes to which 60 μL of Solution C1 has been added using sterile scissors.
9. Extend bead-beating length by 10 min [6].

3.1.1 Oral Microbiome Swabbing Gums, Cheeks, Tongue, Tooth Surface, or Throat

Different oral microbiomes can be obtained by swabbing the gums, inside of cheeks, tongue, tooth surface, or throat, respectively [7]. The microbes obtained by swabbing these inner surfaces contain not only microbes found in the tongue but also those that may adhere to the epithelial cell surfaces.

1. The swab can be any type of sterile cotton-tip collection swab and 20 mL conical tube.
2. Instruct the patient not to eat/drink/smoke/chew gum for at least 30 min prior to giving a sample.
3. Use the sterile swab tip to rub the inside of oral cavity. Site of swabbing will depend on question being asked and area you are interested at. Key is to be consistent in order to replicate results.
4. Store the samples at -80°C until they are ready to be processed.
5. DNA can then be extracted following **steps 5, 8, and 9** above.

Alternate Protocol

1. Instruct the patient not to eat/drink/smoke/chew gum for at least 30 min prior to giving a sample.
2. Have the patient rinse the oral cavity by swishing 10 mL of sterile water for 10 s.
3. Open 50 mL conical tube.
4. Have the patient spit the water into the 50 mL conical tube.
5. Screw the lid tightly onto the collection tube and store the oral wash samples at -80°C until the samples are ready to be processed.
6. DNA can then be extracted following **steps 5 and 7** above.

3.2 Sample Collection and DNA Extraction of Skin Microbiome

1. On the day of sampling, participants should be instructed to bathe in the morning using their choice of personal hygiene products, but to avoid wearing deodorant or antiperspirant.
2. Moisten collection swab in sterile buffer solution.
3. Stretch the area to be swabbed, approximately a 4 cm² area.
4. Rub the swab back and forth applying firm pressure.
5. Swab/swab tip can be placed back into the buffer and store at -80°C until they are ready to be processed.
6. Using the DNeasy PowerWater Kit (Qiagen, Hilden, Germany) according to the manufacturer's instruction DNA can be extracted with the following modification [8]:
7. Swab tips should be incubated with solution C1 in a 65 °C water bath for 15 min prior to bead beating.
8. Bead beating length should be extended to 10 min (with the swabs included).
9. Samples should be eluted in 50 µL of solution PW6.

3.3 Sample Collection and DNA Extraction of Vaginal Microbiome

1. On the day of sampling, participants should be instructed to bathe in the morning using their choice of personal hygiene products, but to avoid wearing deodorant or antiperspirant [9].
2. Participants should be instructed to insert the vaginal swab 1–2 in. into the vagina, twisting the swab to collect material on all the sides of the tip, wipe in several full circles on the vaginal wall and keep swab in vagina for 20 s.
3. Carefully remove the swab and place in the sterile cryovial and store at $-80\text{ }^{\circ}\text{C}$ until they are ready to be processed.
4. DNA will be extracted from swabs using the DNeasy PowerSoil Kit (Qiagen, Hilden, Germany) according to the manufacturer's instruction. DNA can be extracted with the following modification [10]:
5. Break off the cotton tips of the frozen swabs directly into bead tubes to which 60 μL of Solution C1 has been added.
6. Incubate the tubes at $65\text{ }^{\circ}\text{C}$ for 10 min.
7. Shake horizontally at maximum speed for 2 min using the MO BIO vortex adaptor.

3.4 Sample Collection and DNA Extraction of Gut Microbiome

Stool has been used in a number of studies as a proxy for investigating the gut microbial composition [11–14]. However, stool subsampling can result in large variability of gut microbiome data due to different microenvironments harboring various taxa within an individual stool [15]. Homogenizing the entire stool sample in liquid nitrogen and subsampling from this prior to DNA extraction may achieve reduction of intra-sample variability. There are numerous ways to collect stool samples and the approach will largely depend on the model organism used in the study. Protocols below are used for human [16] and mouse samples.

3.4.1 Human Sample Collection

1. Instruct subjects to place a disposable commode specimen container under their toilet seat prior to bowel movement (*see Note 5*).
2. Using the sterile spatula or plastic spoon, put approx. 1 g into a 50 mL conical tube and store at $-80\text{ }^{\circ}\text{C}$ until the samples are ready to be processed.
3. If the samples are to be shipped make sure to use next day delivery.
4. After bowel movement, stool can be wiped off of used bathroom tissue or by directly swabbing into the largest piece of feces; just enough to change the color of the swab should suffice.
5. Place swabs into a sterile 15 mL conical tube and store at $-80\text{ }^{\circ}\text{C}$, or fully submerged in $\geq 95\%$ ethanol, until they are ready to be processed.
6. If the samples are to be shipped make sure to use next day delivery.

3.4.2 *Mouse Feces* *Sample Collection*

In order to maximize the study group replicates, efforts should be made to collect from many individual mice over time. To maintain consistent sampling across all replicates, a daily/weekly collection schedule is recommended.

1. Mice can individually be placed in single-use containers such as Solo-32 ounce poly-coated paper food container with lid.
2. Keep mouse in a container for about 10 min and collect 2–3 fecal pellets into a 1.5 mL vial with sterile forceps. *Be sure to continuously switch to a clean sterile forceps between different mouse sample collections.*
3. Quickly store at -80°C or fully submerged in $\geq 95\%$ ethanol until processing.

3.4.3 *Feces DNA* *Extraction*

1. Extractions should be done according to the manufacturer's instruction with the DNeasy PowerSoil Kit (Qiagen, Hilden, Germany) (*see Note 1*).
2. Fecal samples only require 0.25 g.
3. Add 60 μL of solution C1 into the bead tubes and break the cotton tips of the frozen swabs directly into the tubes.
4. Shake horizontally at maximum speed for 2 min using the MO BIO vortex adaptor.

3.5 *16S rRNA* *Amplicon PCR*

Once the sample DNA is prepared, PCR can be used with unique barcoded primers to amplify the V4 hypervariable region of the 16S rRNA gene to create an amplicon library for Illumina sequencing [4] for individual samples (*see Note 6*).

1. Combine the following into a PCR tube: 13 μL of PCR grade water, 10 μL Platinum Hot Start PCR Master Mix (2 \times) from ThermoFisher, 0.5 μL forward primer (10 μM initial concentration, 0.2 μM final), 0.5 μL reverse primer (10 μM initial concentration, 0.2 μM final), 1 μL template DNA (5–30 ng/ μL), in a final volume of 25 μL .
2. Cycle the PCR reaction as follows:

Step1:	3 min	94 $^{\circ}\text{C}$	} 35 \times
Step2:	45 s	94 $^{\circ}\text{C}$	
Step3:	60 s	50 $^{\circ}\text{C}$	
Step4:	90 s	72 $^{\circ}\text{C}$	
Step5:	10 min	72 $^{\circ}\text{C}$	
Step6:	Hold	4 $^{\circ}\text{C}$	

3. Check amplified PCR products on a 1% agarose gel (w/v) containing DNA gel stain. Expected band size for 515f/806r is approx. 300–350 bp.
4. Quantify PCR product using a NanoDrop, Picogreen, or Qubit assay (*see the manufacturer's protocols*).

3.6 Viral Purification and Nucleic Acid Extraction

1. Sample pretreatment involves homogenizing samples before viral purification. Approaches will depend on sample type. For example, coral tissues require chloroform homogenization of the matrix [17, 18].
2. Chloroform homogenization [19]: 5 mL of chloroform per 40 mL of coral blastate is added and the samples are agitated gently for 1 h at room temperature.
3. Coral blastates are then homogenized at 5000 rpm for 1 min in the Heidolph SilentCrusher™.

3.6.1 Soil, Animal Tissues, or Clinical Samples

1. Place the sample into appropriate buffer, e.g., PBS or saline magnesium buffer [19].
2. Disrupt the sample with an electric homogenizer at approx. 5000 rpm for 30–60 s or use a mortar and pestle.

3.6.2 Concentration of Viral-like Particles

1. Centrifuge at low speed (approx. $2500 \times g$) at room temperature for 5–10 min to pellet tissue, cells, or sediment [19].
2. Transfer the supernatant to a new tube and filter the supernatant to remove microbial and eukaryotic cells and nuclei by sequential passage through 0.45 μm and then a 0.22 μm Whatman filters.

3.6.3 Density Gradients and Ultracentrifugation

1. Pour each density into clear ultracentrifugation tubes (e.g., Beckman Coulter tubes). For phages use 1 mL of: 1.7, 1.5, and 1.35 g/mL CsCl densities, respectively. For other viruses consult the literature for appropriate densities and type of gradient (e.g., CsCl versus sucrose) [19].
2. After the layers are poured, slowly add identical volumes of samples to the top of each gradient.
3. You must exactly balance the tubes prior to centrifugation. Add or remove small drops of the sample to the centrifugation tube until the difference between tubes is less than 1 mg.
4. For phages, centrifuge for 2 h at $\sim 60,000 \times g$ and 4 °C. Depending on the density of the viruses, different speeds and times will apply. To determine the correct centrifuge speed, consult literature sources [17–19].
5. Using a sterile 18-gauge needle, place the needle with the mouth facing upward, just below the appropriate gradient density where the virus is expected to localize.
6. Carefully withdraw the plunger of the syringe and pull the desired volume into the syringe barrel. Transfer the collected virion layers into a sterile 1.5 mL centrifuge tube.

3.6.4 Nucleic Acid Extraction

Isolation of viral nucleic acids can depend on whether the researcher is studying DNA or RNA viruses. If working with DNA viruses from blood or tissue samples, it may be required to DNase I treat the samples to ensure removal of background host DNA. This part

of the procedure can be long and complex depending on the viral community (DNA, RNA, or both) so we recommend that the researcher consult the appropriate literature, but in short this part will involve:

1. QIAamp MinElute Virus Spin Kit can be used to isolate genetic material from DNA viruses and the RNeasy plus Kit (Qiagen) can be used for RNA viruses. Use the kits according to the manufacturer's protocol.
2. Measure DNA/RNA concentration with a Nanodrop or Bioanalyzer.
3. Validate bacterial and eukaryotic DNA and RNA removal using primers for PCR or reverse transcriptase-PCR amplification of the 16S and 18S rDNA genes, respectively (*see* **Notes 2** and **4**).

3.6.5 Random Amplification

cDNA will need to be synthesized if working with RNA (*see* **Note 7**). There are plenty of reverse transcription reagent kits out there. Once you have DNA you can go ahead and amplified using the Genomiphi amplification kit.

1. Mix 1–10 ng of template DNA with 9 μ L of sample buffer, mix and spin down.
2. Denature the sample by heating for 3 min at 95 °C.
3. Cool on ice for 3 min.
4. For each amplification reaction, combine 9 μ L of reaction buffer with 1 μ L of enzyme, mix and place on ice, add to the cooled sample from step before.
5. Incubate the sample for 16–18 h at 30 °C (for a maximum of 18 h, but not shorter than 6 h).
6. Inactivate the enzyme by heating the sample for 10 min at 65 °C.
7. Cool to 4 °C.
8. Clean up the amplified DNA with a modified version of a Qiagen DNeasy Kit.
9. Add 180 μ L of buffer ATL, 200 μ L of buffer AL and 200 μ L of 100% ethanol to the 20 μ L of reaction volume. If the sample is too viscous heat to 55–65 °C.
10. Add to the column provided with the kit, and proceed with the purification steps as specified in the manual.

3.6.6 Sequencing of Libraries

DNA generated from the above steps can be taken into any standard Illumina compatible library preparation kit and sequenced on an Illumina platform sequencing instrument.

4 Notes

1. There are other kits that may be used to isolate DNA, including the ones that are suited for higher throughput of samples such as, QIAamp DNA mini kit and QIAamp DNA stool kit (Qiagen, Valencia, CA). Consult the literature for collection protocols of fecal samples from other organisms.
2. 16S rRNA universal bacterial primers are: 515 forward primer (5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT ATG GTA ATT GTG TGC CAG CMG CCG CGG TAA-3') and 805 reverse primers that also contain unique 12 base pair Golay barcodes (5'-CAA GCA GAA GAC GGC ATA CGA GAT NNNNNNNNNNNN GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC TGG ACT ACH VGG GTW TCT AAT-3') [4].
3. Primer set kits (NEXTflex® 16S V4 Amplicon-Seq Kit 2.0 and NEXTflex® 16S V1-V3 Amplicon-Seq Kit) are commercially available from BioScientific, Austin, TX, USA for amplifying the V1–V3 and V4 variable regions of bacteria rRNA.
4. Universal 18S eukaryotic primers are: 1A (5'-AAC CTG GTT GAT CCT GCC AGT-3') [20] and 564R (5'-GGC ACC AGA CTT GCC CTC-3') [21].
5. Ensure no urine is collected in the sample.
6. Samples should be thawed on ice.
7. Depletion of host DNA extracted RNA samples can be done using DNase treatment in RNeasy plus mini kit (Qiagen Hilden, Germany) following the manufacturer's instructions. Random hexamer oligonucleotides coupled with unique barcodes are used for first-strand cDNA (Life Technologies, Inc.) and second-strand DNA (New England Biolabs Beverly, MA, USA) synthesis [22].

References

1. Nechvatal JM, Ram JL, Basson MD, Namprachan P, Niec SR, Badsha KZ et al (2008) Fecal collection, ambient preservation, and DNA extraction for PCR amplification of bacterial and human markers from human feces. *J Microbiol Methods* 72(2): 124–132
2. Hale VL, Tan CL, Knight R, Amato KR (2015) Effect of preservation method on spider monkey (*Ateles geoffroyi*) fecal microbiota over 8 weeks. *J Microbiol Methods* 113:16–26
3. Gray MA, Pratte ZA, Kellogg CA (2013) Comparison of DNA preservation methods for environmental bacterial community samples. *FEMS Microbiol Ecol* 83(2):468–477
4. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N et al (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* 6(8):1621–1624
5. Torres PJ, Fletcher EM, Gibbons SM, Bouvet M, Doran KS, Kelley ST (2015) Characterization of the salivary microbiome in patients with pancreatic cancer. *PeerJ* 3:e1373
6. Meadow JF, Altrichter AE, Kembel SW, Moriyama M, O'Connor TK, Womack AM

- et al (2014) Bacterial communities on classroom surfaces vary with human contact. *Microbiome* 2(1):7
7. Zaura E, Keijsers BJ, Huse SM, Crielaard W (2009) Defining the healthy “core microbiome” of oral microbial communities. *BMC Microbiol* 9:259
 8. Meadow JF, Bateman AC, Herkert KM, O’Connor TK, Green JL (2013) Significant changes in the skin microbiome mediated by the sport of roller derby. *PeerJ* 1:e53
 9. Forney LJ, Gajer P, Williams CJ, Schneider GM, Koenig SS, McCulle SL et al (2010) Comparison of self-collected and physician-collected vaginal swabs for microbiome analysis. *J Clin Microbiol* 48(5):1741–1748
 10. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N et al (2010) Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc Natl Acad Sci U S A* 107(26):11971–11975
 11. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS et al (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312(5778):1355–1359
 12. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444(7122):1027–1031
 13. Koren O, Goodrich JK, Cullender TC, Spor A, Laitinen K, Bäckhed HK et al (2012) Host remodeling of the gut microbiome and metabolic changes during pregnancy. *Cell* 150(3):470–480
 14. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE et al (2014) Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505(7484):559–563
 15. Gorzelak MA, Gill SK, Tasnim N, Ahmadi-Vand Z, Jay M, Gibson DL (2015) Methods for improving human gut microbiome data by reducing variability through sample processing and storage of stool. *PLoS One* 10(8):e0134802
 16. Kumar R, Eipers P, Little RB, Crowley M, Crossman DK, Lefkowitz EJ et al (2014) Getting started with microbiome analysis: sample acquisition to bioinformatics. *Curr Protoc Hum Genet* 82:18.8.1–18.829
 17. Weynberg KD, Wood-Charlson EM, Suttle CA, Oppen MJ v (2014) Generating viral metagenomes from the coral holobiont. *Front Microbiol* 5:206
 18. Marhaver KL, Edwards RA, Rohwer F (2008) Viral communities associated with healthy and bleaching corals. *Environ Microbiol* 10(9):2277–2286
 19. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. *Nat Protoc* 4(4):470–483
 20. Medlin L, Elwood HJ, Stickel S, Sogin ML (1988) The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene* 71(10):491–499
 21. Wang Y, Tian RM, Gao ZM, Bougouffa S, Qian PY (2014) Optimal eukaryotic 18S and universal 16S/18S ribosomal RNA primers and their application in a study of symbiosis. *PLoS One* 9(3):e90053
 22. Moser LA, Ramirez-Crvajal L, Puri V, Pauszek SJ, Matthews K, Diley KA et al (2016) A universal next-generation sequencing protocol to generate noninfectious barcoded cDNA libraries from high-containment RNA viruses. *mSystems* 1(3):e00039–e00015

Chapter 12

A Bloody Primer: Analysis of RNA-Seq from Tissue Admixtures

Casey P. Shannon, Chen Xi Yang, and Scott J. Tebbutt

Abstract

RNA sequencing is a powerful technology that allows for unbiased profiling of the entire transcriptome. The analysis of transcriptome profiles from heterogeneous tissues, cell admixtures with relative proportions that can vary several fold across samples, poses a significant challenge. Blood is perhaps the most egregious example. Here, we describe in detail a computational pipeline for RNA-Seq data preparation and statistical analysis, with development of a means of estimating the cell type composition of blood samples from their bulk RNA-Seq profiles. We also illustrate the importance of adjusting for the potential confounding effect of cellular heterogeneity in the context of statistical inference in a whole blood RNA-Seq dataset.

Key words RNA-Seq, Transcriptomics, Whole blood, Cellular heterogeneity, Cell type-specific deconvolution

1 Introduction

Genome-wide transcript profiling by RNA-Seq is a powerful technology that allows for unbiased profiling of the entire transcriptome and will soon replace microarrays. Unlike arrays, RNA-Seq does not rely on transcript-specific probes for detection. It is therefore able to quantify novel transcripts, or detect transcript modifications, such as gene fusions, single-nucleotide variants, and small insertions and deletions, that arrays are not able to detect. Moreover, RNA-Seq quantifies discrete, digital sequencing read counts, which are not affected by background (low end) or signal saturation (high end), resulting in a greater dynamic range. Finally, sequencing coverage depth can easily be increased to allow for the detection of rare transcripts, down to the single transcript per cell level.

Nevertheless, RNA-Seq remains susceptible to some of the same analytical pitfalls as microarrays. The analysis of transcriptome profiles from heterogeneous tissues, for example, is a significant challenge. Genome-wide transcript profiling studies are often

carried out using tissue samples, but the conclusions drawn from them are about the behavior of *average* cells. The aggregate transcript abundance across a large population of cells is used as a proxy for studying the behavior of individual cells (though single-cell RNA-Seq is growing in popularity). This is reasonable if the tissue under study is largely composed of a single-cell type. In practice, however, this assumption is often violated. Tissues are typically cell admixtures with relative proportions that can vary several folds across samples.

Perhaps the most egregious example is blood. Blood is often the tissue specimen of choice for large-scale studies of human biology and disease. It is easily collected and provides a global picture of the status of the immune system. It is, however, a complex tissue, composed of many different cell types, each with highly specialized function. The sensitivity and interpretability of analyses carried out in blood are significantly affected by its dynamic heterogeneity. Quantifying this heterogeneity is important, allowing us to account for its effect or to better model interactions that may be present between the abundance of certain transcripts, some cell types, and the indication under investigation. Accurate enumeration of the many component cell types that make up peripheral whole blood can be costly, however, and further complicates the sample collection process. Often, such quantification is an afterthought. When complete enumeration of the blood is not available, its composition may be estimated, at least to some degree of granularity, directly from its transcriptomic profile.

In this chapter, we introduce various approaches for doing this and illustrate the importance of accounting for the cellular heterogeneity present in the data. We use this admittedly more advanced topic to introduce RNA-Seq analysis from first principles. First, we present a simple pipeline for going from the raw files produced by the sequencing instrument to summarized read counts suitable for analysis. Next, we take the resulting read count data into the R statistical computing environment, implement a simple normalization scheme, and develop a means of estimating the composition of mixed blood samples from their RNA-Seq profiles. We do this using freely available data—RNA-Seq reads obtained from some of the component cell types of peripheral whole blood—to identify genes that exhibit distinct patterns of expression across these cell types, termed marker genes. Finally, we use a second publicly available whole blood RNA-Seq dataset to illustrate the importance of adjusting for the potential confounding effect of cellular heterogeneity in the context of statistical inference. We hope that this treatment will be of interest to RNA-Seq beginners and experts alike.

2 Data Processing Overview

Transcriptome profiling by RNA-Seq results in millions of short reads derived from fragments of the sample RNA. The reads produced by such experiments can be used for a number of analyses, including transcript quantification, reference-based gene annotation, and de novo assembly. Here, we focus on transcript quantification for differential expression analysis. This is generally a two-step process: reads are first aligned to a reference genome sequence or set of reference transcript sequences, and gene (or isoform) abundances are then estimated from these aligned reads.

Transcript-level alignment is simpler than genome-level alignment, particularly when processing data from eukaryotic samples. Intron-spanning reads or reads that extend into the poly(A) tails can be challenging to align at the genome level, though specialized tools do exist. Transcript-level alignment is also amenable to quantification in species where a full reference genome assembly is not available, and can be significantly faster, since the transcriptome is typically an order of magnitude smaller than the genome. On the other hand, we potentially miss out on intronic or intergenic features relevant to the biological process under study. Since both approaches are potentially of interest, we have ensured that the collection of software described below is flexible enough to accommodate both transcriptome- and genome-level alignments with minimal modification.

In the following sections, we describe a data processing pipeline for RNA-Seq. This pipeline takes as input the reads produced by an RNA-Seq experiment, in the form of FASTQ files, and outputs both gene- and isoform-level summarized read counts as flat text files, ready for analysis. We first introduce the workflow management tool that powers the pipeline in the section entitled *Workflow Management with Snakemake*. Next, we carry out quality control by looking at a variety of metrics using the FastQC [1] program. This is described in Subheading 4. Reads are then aligned to a set of reference transcript sequences derived from the reference genome sequence. We outline how to do this using the STAR [2] (Spliced Transcripts Alignment to a Reference) program in Subheading 5.1. If a reference genome sequence is not available for the organism you are studying, *de novo* assembly will be required. This procedure is outside the scope of the current chapter, however. Finally, aligned reads are summarized to read counts using RSEM [3] (RNA-Seq by Expectation-Maximization). Details are provided in Subheading 5.2. The resulting summarized read counts are then used for data analysis using the R statistical computing environment in the section entitled *Analysis in R*.

3 Workflow Management with Snakemake

Preparing RNA-Seq data for the analysis is a complex process. In the above section, we outlined an analysis pipeline dependent on a number of distinct software programs. Each of these programs has its own set of parameters, differing inputs and outputs and, because of the large file sizes involved, the entire process is relatively long running. The installation and management of all the moving parts in this data processing pipeline is tedious. Ideally, we would like for the data generating process to be automated, self-documenting, auditable, and reproducible. To achieve this, we leverage the Conda [4] package and environment management system, and a workflow management program called Snakemake [5].

Conda is a cross-platform package and environment management system that greatly simplifies the installation of the necessary software, allowing regular users to install arbitrary versions of many programs, from maintained lists of pre-compiled software. The system creates software environments that can easily be initialized and shut down, allowing users to rapidly switch between configurations. We will install and use the bare bones Miniconda version of the Conda software to handle procurement for our RNA-Seq data processing pipeline.

The pipeline itself is implemented as a Snakemake workflow. A workflow is composed of a series of rules. Each rule specifies input and output files, along with the command or script necessary to go from one to the other. Rules are written using a simple syntax and can invoke shell commands, compiled programs, and Python or R scripts. Dependencies between the rules are determined automatically, creating a directed acyclic graph of jobs that can be parallelized across cores, or even compute cluster nodes. The workflow can be suspended and resumed at any time, without having to start over. In addition, Snakemake will only update output file(s) if one or more of the input file(s) are more recent. Rules are written to a Snakefile, which serves as a record of the data generating process, including programs and parameters used. Snakefiles can be versioned, shared, and re-used. Parameters that are often modified by users may be stored in a separate configuration file, written using a simple markup language, allowing for further flexibility.

We start by cloning the workflow repository from Github. This repository can effectively act as a project template for each new RNA-Seq analysis.

```
# shell  
  
# clone the repository  
git clone https://github.com/PROOF-centre/rnaseq_workflow.git  
  
# move into the repository
```



```
cd rnaseq_workflow
# inspect its contents
ls -la
```

The folder contains a *Snakefile*, where the various workflow rules are defined, a *config.yaml* file, where parameters for the various rules are stipulated, and a *requirements.yaml* file, which lists (and can be used to install) the software needed in order to execute the workflow. We will now briefly describe each rule in this workflow, before finally assembling the complete *Snakefile* in the section entitled *All together now!*

4 Quality Control with FastQC

In the first step of our data processing pipeline, FASTQ files are subjected to quality control using the FastQC program. FastQC performs simple quality control checks on raw sequence data coming from high-throughput sequencing pipelines. It provides a number of analyses that can help determine whether your data has any problems that you should be aware of before doing any further analysis. Guidance on interpretation of the various diagnostic plots is beyond the scope of this chapter, but useful discussion can be found on [FastQC](#) website.

Our Snakemake rule takes each (compressed) FASTQ file in turn and passes it to the FastQC program, which produces an HTML-formatted report. Running time will depend on the total number of reads (approximately 10 min/FASTQ file for 50M reads).

4.1 Snakemake Rule

```
rule fastQC:
    input:
        lambda wildcards: config["samples"][wildcards.
sample]
    output:
        "fastQC/{sample}/"
    benchmark:
        "benchmarks/fastQC/{sample}.benchmark.txt"
    threads: 8
    params:
        outDir = "fastQC/{sample}/"
    shell:
        "zcat -c {input} | fastqc -t {threads} --out-
dir {params.outDir} stdin"
```

5 Quantifying Transcript Abundance

Next, we quantify transcript abundance in each sample using the RSEM and STAR programs. This is a two-step process: reads are first

aligned to a reference genome sequence by STAR and then summarized to counts, at both the gene and isoform levels, by RSEM. For convenience, RSEM includes a script `--- rsem-calculate-expression ---` that performs read alignment using the parameters recommended by the Encyclopedia of DNA Elements (ENCODE) consortium, and carries out read summarization in a single step. The second step of our data processing pipeline invokes this script to align the reads from each FASTQ file (or, in the case of paired-end runs, pair of FASTQ files) and summarize them to gene and isoform counts.

5.1 Read Alignment

Reads are first aligned to the reference genome sequence by the STAR aligner. STAR is a multi-threaded, high-performance alignment program, purpose-built for RNA-Seq. In particular, it greatly improves alignment accuracy for intron-spanning reads. Here, we primarily favor STAR over alternatives, such as bowtie, for its flexibility and speed. It achieves a very high alignment rate by using a specialized in-memory reference genome index. This index can be quite large: if working with the human genome, your system should have >30 GB of random access memory (RAM).

5.1.1 Generating the STAR Index

Before we can process any FASTQ files, we must first build this index, so that it may subsequently be loaded into memory prior to carrying out any alignments. This only needs to be done once before reads can be aligned to a given reference genome sequence. To do this, we will need:

- The reference genome sequence in a FASTA-formatted file.
- A gene annotation file in GTF format.

The most current human reference genome and transcriptome sequences, and corresponding annotation files, can be obtained from the [GENCODE](#) website. Below, we define a pair of Snakemake rules to download these files from the Sanger Institute FTP servers to the `~/genome/` directory on our system. This location is specified by the `downloadDir` parameter in the `config.yaml` file and can be modified by the user.

1. Snakemake Rule

```
rule downloadGTF:
    output:
        expand("{downloadDir}gtf/gencode.v25.annotation.gtf", downloadDir = config["downloadDir"])
    params:
        url = "ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_25/gencode.v25.annotation.gtf.gz",
        downloadDir = config["downloadDir"]
    shell:
        "wget -O - {params.url} | gunzip -c > {output}"
```

2. Snakemake Rule

```
rule downloadFASTA:
    output:
        expand("{downloadDir}fasta/GRCh38.
p7.genome.fa", downloadDir = config["downloadDir"])
    params:
        url = "ftp://ftp.sanger.ac.uk/pub/gencode/
Gencode_human/release_25/GRCh38.p7.genome.fa.gz",
        downloadDir = config["downloadDir"]
    shell:
        "wget -O - {params.url} | gunzip -c > {output}"
```

Next, we invoke the *rsem-prepare-reference* RSEM script, which will call STAR and build a suitable reference genome index.

3. Snakemake Rule

```
rule generateRSEMIndex:
    input:
        gtf = expand("{downloadDir}gtf/gen-
code.v25.annotation.gtf", downloadDir =
config["downloadDir"]),
        fasta = expand("{downloadDir}fasta/GRCh38.
p7.genome.fa", downloadDir = config["downloadDir"])
    output:
        expand("{downloadDir}index/rsem/", down-
loadDir = config["downloadDir"])
    shell:
        "rsem-prepare-reference \
-p 8 \
--star \
--gtf {input.gtf} \
{input.fasta} \
{output}GRCh38"
```

5.1.2 Aligning Reads

Once a suitable genome index has been created, we can begin aligning reads. STAR will take either a single, or a pair of (compressed) FASTQ files for each sample, take each read or read-pair (sometimes referred to as a fragment) in turn, and find the optimal alignment along the reference genome sequence. Reads (or fragments), along with location of the optimal alignment, are returned and written to a new file, called a BAM file. This is the binary version of a SAM file, a tab-delimited text file standard for sequence alignment data.

5.2 Read Summarization

Once all alignments are completed, *rsem-calculate-expression* invokes the RSEM program to produce two tab-delimited text files containing the summarized read counts at the gene and isoform levels.

Our Snakemake rule passes single, or pairs of (compressed) FASTQ files to STAR (via *rsem-calculate-expression*) for alignment using the genome index we created above. Conveniently, STAR

can read from compressed FASTQ files directly (specified using `--star-gzipped-read-file`). Since our focus here is on analysis of gene and isoform read counts, we choose to discard the large BAM files following quantification (with `--no-bam-output`), but note that they may be of interest for other types of analysis.

1. Snakemake Rule

```
rule RSEM:
    input:
        fastqs = lambda wildcards: config["samples"]
        [wildcards.sample],
        ref = expand("{downloadDir}index/rsem/",
        downloadDir = config["downloadDir"])
    output:
        "rsem/{sample}.genes.results",
        "rsem/{sample}.isoforms.results"
    benchmark:
        "benchmarks/rsem/{sample}.benchmark.txt"
    threads: 8
    params:
        readsType = config["readsType"],
        id = "{sample}"
    shell:
        "rsem-calculate-expression \
        {params.readsType} \
        --star \
        --star-gzipped-read-file \
        -p {threads} \
        --no-bam-output \
        {input.fastqs} \
        {input.ref}GRCh38 \
        {params.id} && \
        mv -v *.genes.results *.isoforms.results
        *.stat/ -t rsem/"
```

6 All Together Now!

We have now described all the rules required to go from raw RNA-Seq reads contained in FASTQ files to summarized read counts for a set of annotated genomic features. In this section, we pull together all these rules into a single *Snakefile* that both powers, and serves as a record of, our data processing pipeline, and introduce the *config.yaml* file—a mechanism that allows us to customize this pipeline for different analytical requirements.

6.1 The Snakefile

```
configfile: "config.yaml"
rule all:
    input:
        expand("fastQC/{sample}/", sample =
        config["samples"]),
```

```

        expand("rsem/{sample}.genes.results", sample =
config["samples"]),
        expand("rsem/{sample}.isoforms.results", sample
= config["samples"])
    shell:
        "conda env export > environment.yaml"

rule downloadGTF:
    output:
        expand("{downloadDir}gtf/gencode.v25.annota-
tion.gtf", downloadDir = config["downloadDir"])
    params:
        url = "ftp://ftp.sanger.ac.uk/pub/gencode/
Gencode_human/release_25/gencode.v25.annotation.gtf.gz",
        downloadDir = config["downloadDir"]
    shell:
        "wget -O - {params.url} | gunzip -c > {output}"

rule downloadFASTA:
    output:
        expand("{downloadDir}fasta/GRCh38.p7.genome.
fa", downloadDir = config["downloadDir"])
    params:
        url = "ftp://ftp.sanger.ac.uk/pub/gencode/
Gencode_human/release_25/GRCh38.p7.genome.fa.gz",
        downloadDir = config["downloadDir"]
    shell:
        "wget -O - {params.url} | gunzip -c > {output}"

rule generateRSEMIndex:
    input:
        gtf = expand("{downloadDir}gtf/gencode.v25.an-
notation.gtf", downloadDir = config["downloadDir"]),
        fasta = expand("{downloadDir}fasta/GRCh38.
p7.genome.fa", downloadDir = config["downloadDir"])
    output:
        expand("{downloadDir}index/rsem/", downloadDir
= config["downloadDir"])
    shell:
        "rsem-prepare-reference \
        -p 8 \
        --star \
        --gtf {input.gtf} \
        {input.fasta} \
        {output}GRCh38"

rule fastQC:
    input:
        lambda wildcards: config["samples"][wildcards.
sample]
    output:
        "fastQC/{sample}/"
    benchmark:

```

```

        "benchmarks/fastQC/{sample}.benchmark.txt"
    threads: 8
    params:
        outDir = "fastQC/{sample}/"
    shell:
        "zcat -c {input} | fastqc -t {threads} --out-
dir {params.outDir} stdin"

rule RSEM:
    input:
        fastqs = lambda wildcards: config["samples"]
[wildcards.sample],
        ref = expand("{downloadDir}index/rsem/", down-
loadDir = config["downloadDir"])
    output:
        "rsem/{sample}.genes.results",
        "rsem/{sample}.isoforms.results"
    benchmark:
        "benchmarks/rsem/{sample}.benchmark.txt"
    threads: 8
    params:
        readsType = config["readsType"],
        id = "{sample}"
    shell:
        "rsem-calculate-expression \
{params.readsType} \
--star \
--star-gzipped-read-file \
-p {threads} \
--no-bam-output \
{input.fastqs} \
{input.ref}GRCh38 \
{params.id} && \
mv -v *.genes.results *.isoforms.results
*.stat/ -t rsem/"

```

6.2 The config.yaml File

Some parameters, like directory locations or number of threads used, are specific to the system on which the pipeline will be run. Others, like whether to run STAR or RSEM in paired-end mode, may need to be modified for every new analysis. We recommend using the *Snakefile* as a static description of the data processing pipeline. Each new project can be initialized with a copy of the static *Snakefile*, while the *config.yaml* file can be modified with project-specific parameters. We divide the *config.yaml* file into two parts: the first part contains the parameters passed to RSEM and the STAR aligner (e.g., the path to the genome reference folder or whether the reads are paired-end reads or not), while the second part lists the samples to be processed, including sample IDs and paths to their corresponding FASTQ file(s).

6.3 Running the Pipeline

6.3.1 Installing the Necessary Software

A number of programs need to be installed and placed on the system's PATH before we can run our RNA-Seq data processing workflow. We describe below how to set up a suitable environment on a Linux system, as a regular user without administrative privileges, using Miniconda.

1. At the command line, install Miniconda.

```
# shell

# download the install script
wget https://repo.continuum.io/miniconda/
Miniconda3-latest-Linux-x86_64.sh

# run it
bash Miniconda3-latest-Linux-x86_64.sh

# cleanup
rm Miniconda3-latest-Linux-x86_64.sh
```

2. Create a named Miniconda environment with the required software.

```
# shell
conda env create -n "rnaseq" -f "requirements.yaml"
```

3. Activate the newly created workspace.

```
# shell
source activate "rnaseq"
```

All software specified in the *requirements.yaml* file should now be available to use at the command line.

6.3.2 A Brief Note on Storage Requirements

The raw read data from an RNA-Seq experiment are delivered in the form of FASTQ files. These are large files and usually compressed. File size is a factor of read length and depth of coverage. FASTQ file sizes can be computed as $(\text{Read Length} \times 2 + 50) \times (\text{Number of Reads})$, where the $2\times$ factor accounts for the read and accompanying quality scores, and 50 bytes are added for the identifier. Paired end runs carry an additional $2\times$ multiplier. FASTQ files are typically compressed, and general use compression programs, such as GZIP, achieve a worst case compression ratio of 0.3. Thus, if our experiment was a paired end run, with a target of 50M reads per sample, storage requirements for the raw FASTQ files could be estimated as:

$$\begin{aligned}
 \text{Single Reads} &= 150_{\text{bytes/read}} + 150_{\text{bytes Phred scores}} + 50_{\text{bytes header}} \\
 \text{Reads / Sample} & \qquad \qquad \qquad \qquad \qquad \qquad \times 50,000,000 \\
 \text{Compression Factor} & \qquad \qquad \qquad \qquad \qquad \qquad \times 1/3 \\
 \text{Storage (Single end)} & \qquad \qquad \qquad \qquad \qquad \qquad = 6 \text{ GB / Sample (1 FASTQ File)} \\
 \text{Reads / Fragment} & \qquad \qquad \qquad \qquad \qquad \qquad \times 2 \\
 \text{Storage (Paired end)} & \qquad \qquad \qquad \qquad \qquad \qquad = 12 \text{ GB / Sample (2 FASTQ Files)}
 \end{aligned}$$

Actual storage requirements to process the data and carry out analysis may be higher, however, since many intermediate files, particularly aligned read files (BAM), need to be generated and stored, at least temporarily. We offer the following rule of thumb: total storage requirements for the experiment = $2.5 \times$ FASTQ storage requirement.

4. Create a *fastq* folder under the project directory.

```
# shell
# create a new directory for the project
mkdir fastq
```

5. Move the FASTQ files from your experiment into the newly created directory.

```
# shell
# move all FASTQ files into it, e.g.
mv /path/to/experiment/*.fastq.gz fastq/.
```

6.3.3 Invoking Snakemake

Before running the pipeline, we can do a dry run.

6. Preview the pipeline execution with the *-n* and *-p* flags.

```
# shell
snakemake -np
```

Finally, we run the pipeline.

7. Run the pipeline, utilizing multiple cores with the *--cores* flag.

```
# shell
snakemake -p --cores 8
```

6.3.4 Folder Structure

Here is what our home directory should look like after running the pipeline.

```
|-[Miniconda3]
|  |-This is your working environment with all the
|  |tools installed inside.
|-[genome]
|  |-[gtf]
|  |  |-gencode.v25.annotation.gtf
|  |-[fasta]
|  |  |-GRCh38.p7.genome.fa
|  |-[index]
|  |  |-[rsem]
|  |    |-This folder contains the RSEM genome
|  |    reference files.
|-[rnaseq_workflow]
|  |-[requirements.yaml]
|  |-[Snakefile]
|  |-[config.yaml]
|  |-[fastq]
|  |  |-sample1_1.fastq.gz
```



```

|-sample1_2.fastq.gz
|-sample2_1.fastq.gz
|-sample2_2.fastq.gz
...
|-[fastQC]
  |-[sample1]
    |-stdin_fastqc.html
    |-stdin_fastqc.zip
  |-[samples2]
    |-stdin_fastqc.html
    |-stdin_fastqc.zip
  ...
|-[rsem]
  |-sample1.genes.results
  |-sample1.isoforms.results
  |-[sample1.stat]
  |-sample2.genes.results
  |-sample2.isoforms.results
  |-[sample2.stat]
  ...
|-[benchmarks]
  |-[fastQC]
  |-[rsem]

```

When invoked for the first time, the pipeline will call the *downloadGTF*, *downloadFASTA*, and *generateRSEMIndex* rules to download the GTF annotation and reference genome sequence FASTA files, and generate the necessary STAR genome index. If the genome index already exists, however, the pipeline will skip these rules and immediately start to process the FASTQ files specified in the *config.yaml* file.

7 Analysis in R

We are now ready to carry out analysis of the resulting summarized read counts in R. Recall that, at the beginning of this chapter, we stated that it is important to adjust for cellular composition when analyzing genome-wide transcript abundance data generated from tissues that are cellular admixtures, such as blood. In the following section, we hope to illustrate this point by:

- Demonstrating an approach for obtaining estimates of the cellular composition of admixed samples directly from their RNA-Seq profiles.
- Comparing the results of carrying out differential gene expression analysis using a simple linear model-based approach, with or without correcting for this confounding effect using the cellular composition estimates.

The approach we will use to estimate cellular composition is described in detail in Chikina and others [6]. Briefly, we will estimate

a set of surrogate proportion variables (SPVs) by cross-referencing putative marker genes with the data correlation structure. An implementation of this approach is available in the *CellCODE* R package.

We will identify marker genes for use with the *CellCODE* approach by applying different selection strategies to publicly available gene expression profiles obtained from many leukocyte subpopulations, isolated from peripheral whole blood in healthy individuals ([GSE60424](#)). We will then use the resulting marker gene sets to derive SPVs in a small collection of whole blood RNA-Seq gene expression profiles ([GSE53655](#) [7]) where cellular composition is known, allowing us to determine how well we did. For simplicity, the raw reads from these two experiments were obtained from GEO and processed using the pipeline described above. The resulting summarized read counts are made available in the workflow repository (in the *EGEOD60424* and *EGEOD53655* folders, respectively) as compressed, tab-delimited files.

7.1 Preparing the R Environment

The Conda environment we created and activated above installed R unto the system. The base R installation includes many useful functions, but before we start, we need to install a few additional packages to tackle the proposed analysis:

1. *tidyr* and *dplyr*: our preferred functions for general data manipulation.
2. *purrr*: enables use of functional programming idioms with the `%>%` operator popularized by *dplyr*.
3. *readr*: read (compressed) flat text file into the R environment fast.
4. *ggplot2*: a powerful, high-level, graphics package for R.
5. *matrixStats*: fast implementation of row and column-wise summary statistics.
6. *limma*: differential gene expression using flexible linear models.
7. *glmnet*: fitting generalized linear models *via* penalized maximum likelihood.
8. *CellCODE*: estimate cellular composition of mixed tissue samples from their gene expression using a latent variable approach. First, we start R.

```
# shell
R
```

From here onward, we will be working at the R command line. The code below installs the necessary packages and their dependencies.

```
# include Bioconductor repositories
setRepositories(ind = 1:2)
```

```
install.packages(c('tidyverse', 'cowplot', 'UpSetR',
'matrixStats', 'reshape2', 'limma', 'edgeR', 'glmnet'),
dependencies = T)
```

In addition, we will be using the *CellCODE* package, which, as of this writing, is not available on CRAN or Bioconductor. The code below installs it from the source files provided by the method's authors.

```
# CellCODE depends on gplots and sva, install these
first
install.packages(c('gplots', 'sva'))

# next, install CellCODE from source
url <- 'http://www.pitt.edu/~mchikina/CellCODE/
CellCODE_0.99.0.tar.gz'
install.packages(url, repos = NULL)
```

Finally, we load some of these packages into our environment.

```
# utilities
library(tidyr)
library(dplyr)
library(purrr)

# plotting
library(ggplot2)
library(cowplot)

# stats
library(limma)
library(glmnet)
```

In the next section, we will read the summarized read counts from [GSE60424](#) into R, perform a simple scale normalization of the data and select marker genes, by carrying out differential expression analysis using the *limma* R package [8], using the approach favoured by Chikina and others [6], or using a penalized regression approach *via* the *glmnet* R package [9].

7.2 Loading Read Counts into R

First, we read the summarized counts and associated metadata into R, and tidy up a bit.

```
# list summarized read count files
tsvs <- list.files(path = 'EGEOD60424', pattern =
'*.genes.results.gz',
recursive = T, full.names = T)

# pretty names
names(tsvs) <- gsub( '^EGEOD60424/(.+)\.genes.',
'\1', tsvs)

# helper to convert to matrix with rownames
make_matrix <- function(x) {
  mat <- as.matrix(x[, -1])
  rownames(mat) <- first(x)
  mat
}
```

```

# map over files, read-in
geo <- tsvs %>%
  map(readr::read_tsv) %>%
  map(select, gene_id, expected_count) %>%
  map(make_matrix) %>%
  reduce(cbind)
colnames(geo) <- names(tsvs)
rm(tsvs)
# read-in metadata
meta <- readr::read_csv('EGEOD60424/meta_EGEOD60424.
csv.gz')

# subset to meaningful columns, keep healthy samples,
drop wholeblood
meta <- meta %>%
  select(sample_id = 1, grp = Disease_status, cell =
Cell_type) %>%
  mutate(geo = '60424',
         cell = make.names(tolower(cell))) %>%
  filter(grp == 'Healthy Control', cell != 'whole.
blood')

# enforce that order of samples is the same
geo <- geo[ , meta$sample_id]

# rename samples to correspond to cell type
colnames(geo) <- meta$cell

```

The result is a matrix of read counts, with rows for genes and columns for samples. *limma*'s *voom* function, which we will need to invoke before carrying out differential expression analysis, requires that rows (genes) with zero or very low counts are removed, so we do that next. We choose to remove any row where fewer than three samples (corresponding to the size of our smallest class of samples) have greater than 20 counts.

```

# filter on low counts
counts_above_20 <- apply(geo, 1, function(x) sum(x >
20))
geo <- geo[counts_above_20 >= 3, ]

```

7.3 Applying Scale Normalization

It is usual to apply scale normalization to RNA-Seq read counts. We use a simple total count normalization.

```

# counts per million mapped reads
geo_norm <- t(t(geo) * 1e6/colSums(geo))

```

Comparing the resulting matrices by principal components analysis (PCA) shows that this approach works reasonably well, eliminating most of the variation that was present within groups of samples (compare Figs. 1 and 2).

```

# raw counts
pca <- geo %>%
  t() %>%

```

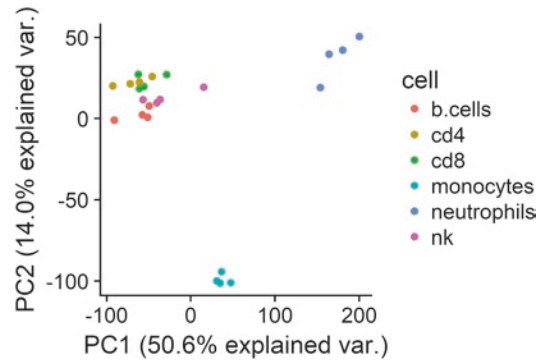


Fig. 1 PCA—raw counts. Principal component analysis applied to the non-normalized, log-transformed count data. Most of the variation in the data can be attributed to large differences between the cell-types. Within cell-types, biological replicates cluster together, but not particularly well when using non-normalized data

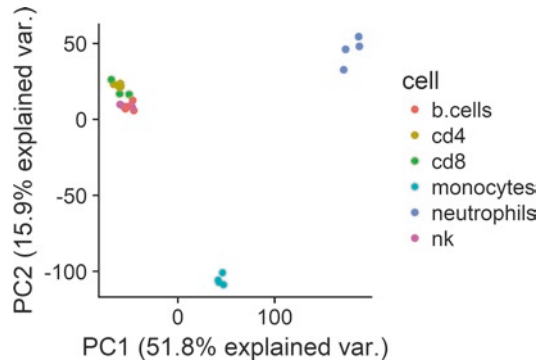


Fig. 2 PCA—normalized counts. Principal component analysis applied to the scale-normalized, log-transformed count data. Again, most of the variation in the data can be attributed to large differences between the cell-types. Within cell types, biological replicates cluster together more closely when using the scale-normalized data. Identifying marker genes using various methods

```

asinh() %>%
  prcomp(center = T, scale. = T)
# normalized - counts per million mapped reads
pca_norm <- geo_norm %>%
  t() %>%
  asinh() %>%
  prcomp(center = T, scale. = T)
```

7.4 See Fig. 2

It is clear from these plots that finding genes able to distinguish between B cells, CD4+, CD8+ T cells and NK cells may be significantly more challenging than identifying genes distinctive of monocytes or neutrophils.

7.4.1 Using *limma*

We first attempt this using *limma*. The code below is equivalent to a one-way ANOVA for each gene except that the residual mean squares have been moderated between genes.

```
# design matrix
design <- model.matrix(~ 0 + cell, data = meta)

# create list of cell types
colnames(design) <- meta$cell %>% unique() %>% sort()

# create list of all pairwise comparisons between cell
types
cells_combos <- colnames(design) %>%
  combn(2, simplify = F) %>%
  map(paste, collapse = '-') %>%
  unlist()

# create coefficients for all pairwise comparisons
contrast_matrix <- makeContrasts(contrasts = cells_
  combos,
                                levels =
colnames(design))

# finally, fit the model
fit_contrasts <- geo_norm %>%
  voom(design) %>%
  lmFit(design) %>%
  eBayes() %>%
  contrasts.fit(contrast_matrix) %>%
  eBayes()

# get the most variable genes across cell types
markers <- fit_contrasts %>%
  topTable(number = 60) %>%
  rownames()
```

This looks reasonable (Fig. 3), but we can do better. As expected, we see a clear distinction between myeloid and lymphoid cell lineages, but less separation between different lymphoid cell types.

7.4.2 Using
the *CellCODE* Method

Next, we try the approach recommended by Chikina and others. This approach is similar to the above, but marker genes must also pass a minimum fold-change cutoff criteria. An implementation is available in the *CellCODE* package.

```
# let CellCODE pick marker genes
markers_cellcode <- geo_norm %>%
  reshape2::melt() %>%
  reshape2::acast(Var1 ~ Var2, mean) %>%
  CellCODE::tagData(., cutoff = 2, max = 10)
```

Again, we see a clear distinction between myeloid and lymphoid cell lineages (Fig. 4), but this time there are significantly more differences between the various T lymphocyte subtypes.

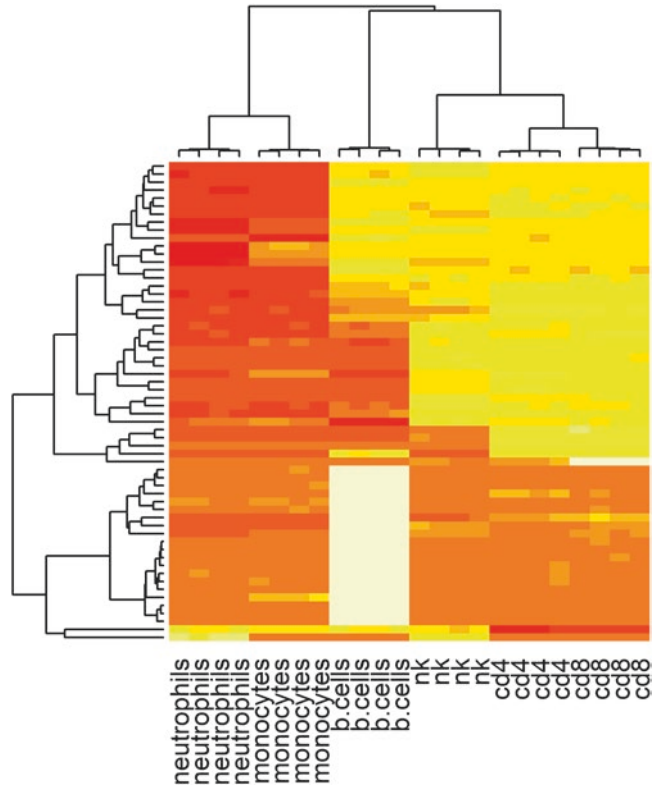


Fig. 3 Clustered heatmap of the top 60 marker genes identified by this approach. The expression of 60 marker genes identified by ANOVA is visualized using a clustered heatmap

7.4.3 Using Penalized Regression with *glmnet*

Conceptually, the identification of marker genes is related to feature selection in the context of classification. *glmnet* is an R package that fits generalized linear models via penalized maximum likelihood. It performs feature selection and can be used for classification or regression problems. We have previously used *glmnet* to derive useful marker genes [10].

The code below uses *glmnet* to identify a minimal subset of genes capable of classifying cell types from their gene expression profiles.

```
# set seed to make cross-validation reproducible
set.seed(123)

# fit model
model <- cv.glmnet(t(geo_norm), colnames(geo_norm),
family = 'multinomial')

# extract selected features
markers_enet <- predict(model, type = 'nonzero', s =
'lambda.1se')
```

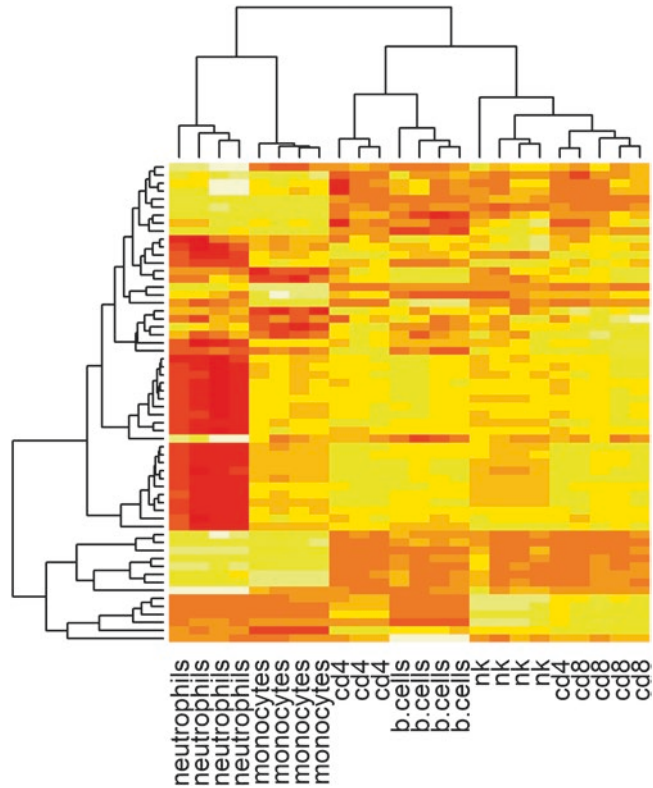


Fig. 4 Clustered heatmap of the 60 marker genes identified by this approach. The expression of 60 marker genes identified by CellCODE is visualized using a clustered heatmap

This looks much better (Fig. 5). The selected set of genes exhibit distinctive patterns of expression across all cell types.

We note that the marker genes identified using these different approaches are almost entirely distinct (Fig. 6).

```
# compare marker gene sets
list(anova = markers,
     cellcode = rownames(markers_cellcode),
     glmnet = markers_enet) %>%
  UpSetR::fromList() %>%
  UpSetR::upset()
```

7.5 Estimating the Cellular Composition of Blood

Next, we will use these marker genes to infer the composition of whole blood samples from their RNA-Seq profiles ([GSE53655](#)) using the *CellCODE* approach.

First, we read in some whole blood RNA-Seq profiles:

```
# blood transcriptome profiles
# list summarized read count files
tsvs <- list.files(path = 'EGEOD53655', pattern =
  '*.genes.results.gz',
```



```

recursive = T, full.names = T)
# pretty names
names(tsvs) <- gsub( '^EGEOD53655/(.)\\.genes.+',
  '\\1', tsvs)
# map over files, read-in
exp <- tsvs %>%
  map(readr::read_tsv) %>%
  map(select, gene_id, expected_count) %>%
  map(make_matrix) %>%
  reduce(cbind)
colnames(exp) <- names(tsvs)
rm(tsvs)
# scale normalization
exp <- t(t(exp) * 1e6/colSums(exp))
# read in metadata
y <- readr::read_tsv('E-GEOD-53655.meta.sdrf.txt')
# remove duplicate columns in geo metadata
y <- y[ , !duplicated(colnames(y))]
# subset to variables of interest
y <- y %>%
  select(run = `Comment [ENA_RUN]`,
         sex = `Characteristics [Sex]`,
         wbc = `Characteristics [wbc 10e9/l]`,
         neu = `Characteristics [neu 10e9/l]`,
         lym = `Characteristics [lym 10e9/l]`,
         mono = `Characteristics [mono 10e9/l]`) %>%
  filter(run %in% colnames(exp)) %>%
  transmute(run = run,
            sex = factor(sex, levels = c('female',
            'male'), labels = c('F', 'M')),
            neu = neu/wbc,
            lym = lym/wbc,
            mono = mono/wbc) %>%
  distinct()
# markers are returned as a list of indices in the
training data by
# predict.cv.glmnet, so order of features must match
exp <- exp[rownames(geo_norm), y$run]

```

7.5.1 Predicting Cell Proportions Using CellCODE

Next, we use the *getAllSPVs* function from the *CellCODE* package, to obtain surrogate proportion variables (SPVs) directly from the whole blood RNA-Seq profiles (Fig. 7).

```

# CellCODE markers
cc <- CellCODE::getAllSPVs(data = exp,
                           grp = y$sex,
                           dataTag = markers_cellcode)
pred_cellcode <- cc %>%
  as.data.frame() %>%
  mutate(run = colnames(exp)) %>%
  gather(markers, spv, -run) %>%
  select(markers, run, spv)

```

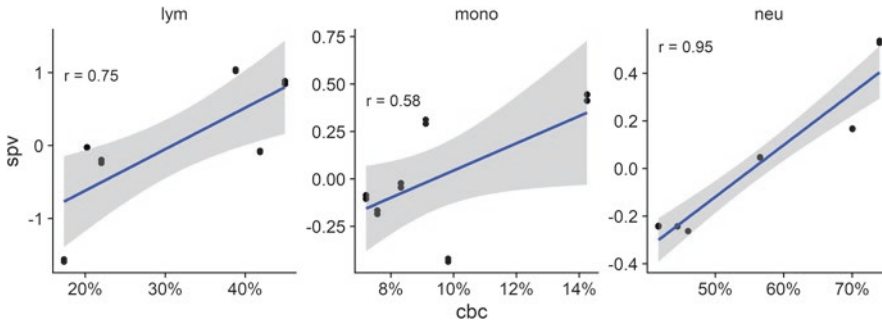


Fig. 7 CellCODE-derived SPVs vs. leukocyte differential counts. Scatterplots comparing the CellCODE-derived surrogate proportion variables (SPVs) to cell percentages obtained from a complete blood count, including leukocyte differential. The coefficient of correlation (Pearson's r) is shown for each cell-type (top-left of each plot), as well as a linear best fit line (blue) with 95% confidence intervals (grey)

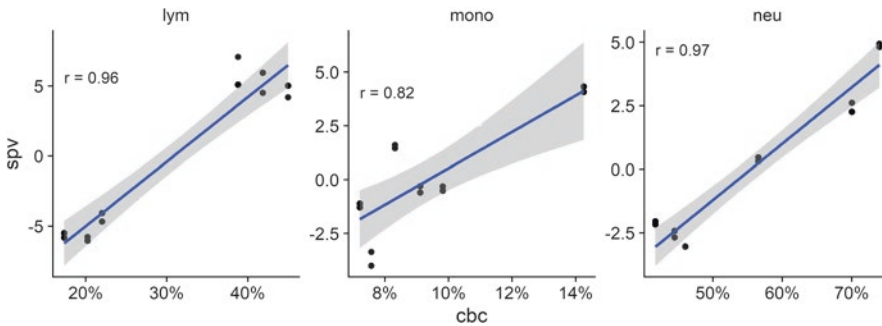


Fig. 8 glmnet-derived SPVs vs. leukocyte differential counts. Scatterplots comparing the glmnet-derived surrogate proportion variables (SPVs) to cell percentages obtained from a complete blood count, including leukocyte differential. The coefficient of correlation (Pearson's r) is shown for each cell-type (top-left of each plot), as well as a linear best fit line (blue) with 95% confidence intervals (grey)

7.5.2 Predicting Cell Proportions Using glmnet-selected Features

We repeat the procedure, this time using the features retained by elastic net as marker genes to derive the SPVs (Fig. 8).

```
# glmnet markers
pred_enet <- model %>%
  predict(type = 'nonzero', s = 'lambda.1se') %>%
  map(unlist, use.names = F) %>%
  map(~ {
    # remove low variance
    x <- t(exp[.x, ])
    x <- x[ , matrixStats::colVars(x) > 0]
    # svd
    pca <- prcomp(x, center = T, scale. = T)
    # extract 1st PC
    pca$x[ , 1]
  }) %>%
  map(as.list) %>%
```

```
map(data.frame) %>%
bind_rows(.id = 'markers') %>%
gather(run, spv, -markers)
```

It is clear that marker gene selection is paramount to deriving useful SPVs. The resulting SPVs can then be used to adjust for the cellular composition of our samples when carrying out differential expression in this dataset.

7.6 Adjusting for Cellular Composition in Linear Models

Finally, we use the estimates we just derived to compare the results of a simple differential gene expression analysis, comparing males to females, with or without adjusting for the potential confounding effect of cellular composition.

Again, we use *limma* to do this. We first fit a linear model that only includes the response variable (*sex*) to the data.

```
# join to predicted cell proportions
y <- pred_enet %>%
  spread(markers, spv) %>%
  right_join(y, by = 'run')

# non-specific filter on variance
var_filter <- function(x) {
  vars <- matrixStats::rowVars(x)
  x[vars >= median(vars), ]
}

# unadjusted
d_unadjusted <- model.matrix(~ sex, data = y)
unadjusted <- exp %>%
  var_filter() %>%
  voom(d_unadjusted) %>%
  lmFit(d_unadjusted) %>%
  eBayes(robust = T, trend = T)
```

Next, we fit a linear model that also includes the proportion of the various component cell types of blood as covariates.

```
# adjusted
d_adjusted <- model.matrix(~b.cells+cd4+cd8+monocytes+
neutrophils+sex, data = y)
adjusted <- exp %>%
  var_filter() %>%
  voom(d_adjusted) %>%
  lmFit(d_adjusted) %>%
  eBayes(robust = T, trend = T)
```

We can compare the results obtained by fitting the two different models by looking at the distribution of the *q*-values (*p*-values adjusted for multiple comparisons using the Benjamini-Hochberg procedure; Fig. 9).

```
# combine into a list
models <- list(unadjusted = unadjusted, adjusted = ad-
justed)
```

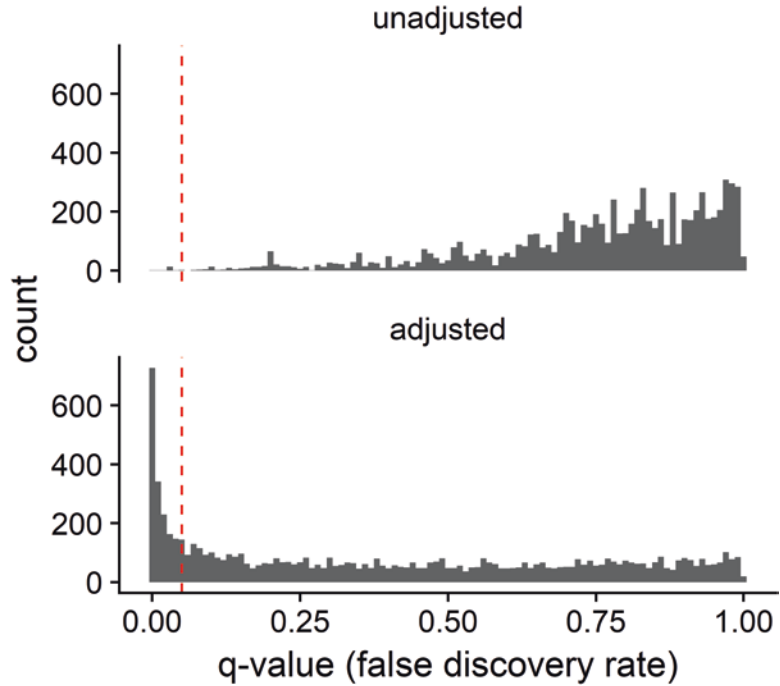


Fig. 9 Distribution of q -values, with and without adjusting for cellular composition. The distribution of q -values (false discovery rate) obtained when identifying differentially abundant genes between males and females is visualized for a linear model including (*bottom*) or not (*top*) cellular proportion estimates as covariates

```
# plot distribution of adjusted p-values
models %>%
  map(topTable, coef = 'sexM', number = Inf) %>%
  map(as.data.frame) %>%
  map(select, adj.P.Val) %>%
  bind_rows(.id = 'Model') %>%
  mutate(Model = factor(Model, levels =
c('unadjusted', 'adjusted'))) %>%
  ggplot(aes(adj.P.Val)) +
  geom_histogram(binwidth = 0.01) +
  geom_vline(xintercept = 0.05, linetype = 'dashed',
colour = 'red') +
  scale_x_continuous(breaks = c(0, 0.1, 0.25, 0.5,
0.75, 1)) +
  facet_wrap(~Model, nrow = 2) +
  theme(strip.background = element_blank()) +
  labs(x = 'q-value (false discovery rate)')
```

Clearly, cellular heterogeneity had a significant effect on our ability to detect differentially expressed genes in this example. Hopefully, this result highlights the importance of incorporating cellular composition when analyzing gene expression data obtained from tissue admixtures.

8 Notes

8.1 Acronyms and Abbreviations

FASTA: A text-based format for storing nucleotide sequences.

FASTQ: A text-based format for storing both nucleotide sequences and corresponding quality scores.

GFF: General feature format. A tab-delimited text file format used for describing genes and other features of DNA, RNA and protein sequences.

GTF: Gene transfer format. A tab-delimited text file format used to hold information about gene structure. A refinement of the general feature format (GFF).

SAM: A text-based format for storing biological sequences aligned to a reference sequence.

BAM: Binary compressed SAM format.

STAR: Spliced Transcripts Alignment to a Reference. A high-performance alignment program purpose-built for processing RNA-Seq reads.

8.2 Online Resources

Code repository:

https://github.com/PROOF-centre/rnaseq_workflow.git

Miniconda

<http://conda.pydata.org/miniconda.html>

Snakemake:

<https://bitbucket.org/snakemake/snakemake/wiki/Home>

Gencode (current release):

<http://www.gencodegenes.org/releases/current.html>

FastQC:

<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

STAR aligner:

<https://github.com/alexdobin/STAR>

RSEM:

<http://deweylab.github.io/RSEM>

References

1. Andrews S. FastQC – a quality control tool for high throughput sequence data. At <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
2. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21
3. Li B, Dewey CN (2011) RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323
4. Conda – package, dependency and environment management for any language. At <http://conda.pydata.org/miniconda.html>
5. Köster J, Rahmann S (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 28:2520–2522
6. Chikina M, Zaslavsky E, Sealfon SC (2015) CellCODE: a robust latent variable approach

- to differential expression analysis for heterogeneous cell populations. *Bioinformatics* 31:1584–1591
7. Shin H, Shannon CP, Fishbane N, Ruan J, Zhou M, Balshaw R, Wilson-McManus JE, Ng RT, McManus BM, Tebbutt SJ (2014) Variation in RNA-Seq transcriptome profiles of peripheral whole blood from healthy individuals with and without globin depletion. *PLoS ONE* 9:e91041
 8. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43:e47–e47
 9. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Statist Soc B* 67:301–320
 10. Shannon CP, Balshaw R, Ng RT, Wilson-McManus JE, Keown P, McMaster R, McManus BM, Landsberg D, Isbel NM, Knoll G, Tebbutt SJ (2014) Two-stage, in silico deconvolution of the lymphocyte compartment of the peripheral whole blood transcriptome in the context of acute kidney allograft rejection. *PLoS ONE* 9:e95224

Next-Generation Sequencing of Genome-Wide CRISPR Screens

Edwin H. Yau and Tariq M. Rana

Abstract

Genome-wide functional genomic screens utilizing the clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9 system have proven to be a powerful tool for systematic genomic perturbation in mammalian cells and provide an alternative to previous screens utilizing RNA interference technology. The wide availability of these libraries through public plasmid repositories as well as the decreasing cost and speed in quantifying these screens using high-throughput next-generation sequencing (NGS) allows for the adoption of the technology in a variety of laboratories interested in diverse biologic questions. Here, we describe the protocol to generate next-generation sequencing libraries from genome-wide CRISPR genomic screens.

Key words CRISPR, Cas9, Genome-wide screen, Genome engineering, GeCKO

1 Introduction

Initially discovered as a prokaryotic adaptive immune system, the bacterial type II clustered regularly interspaced short palindrome repeats (CRISPR) and their associated proteins (Cas9) system have been adapted into mammalian cells to generate a simple and efficient means of generating targeted loss-of-function mutations [1–7]. Cas9 proteins generate precise double-strand breaks at target loci determined by short guide RNAs (sgRNAs). In the absence of a homologous template, these breaks are repaired by the error-prone non-homologous end-joining (NHEJ) mechanism leading to the production of indels. By targeting coding regions, knockout libraries can be generated due to the introduction of indels that lead to frame-shifts. The combination of the simplicity of the CRISPR-Cas9 system, in which the short 20 base pair sgRNA sequence confers the targeting specificity of the Cas9 nuclease, economic oligonucleotide synthesis using pooled microarray synthesis, and the ability to use NGS sequencing for quantifying readout has enabled the use of pooled genome-scale loss-of-function CRISPR screens to systematically interrogate gene function in different biologic processes. Pooled

CRISPR screening has been used to identify genes essential in different genetic contexts [8–10] as well as to identify genes involved in drug or toxin resistance [11–14]. CRISPR screens appear to be more robust in comparison to previous screens utilizing RNA interference technology with short-hairpin RNAs (shRNAs) which were hampered by off-target effects and incomplete genetic knockdown [10].

The most widely adopted system thus far has been the Genome-Scale CRISPR Knock Out (GeCKO) library generated by the Feng Zhang lab and distributed on Addgene and the current protocol is adapted from the methods outlined by this lab which can be found online at <http://genome-engineering.org/gecko/>. The GeCKO library is available targeting either mouse or human coding genes and is available as a 1 plasmid system where sgRNAs and Cas9 protein are encoded on the same plasmid as well as a 2 plasmid system where sgRNAs and Cas9 protein are on separate lentiviral expression plasmids. Each library contains two sublibraries (A and B) with each sublibrary containing 3 sgRNAs targeting each coding gene. A similar library was also generated by the Sabatini and Lander labs targeting human protein coding genes (Addgene #1000000067) as well as targeted sublibraries (Addgene #51043–51048). Second generation libraries built by improving the specificity of sgRNA guides include the Toronto knockout (TKO) library generated by the Moffat lab (Addgene #1000000069) and the Brunello (Addgene #73178 or #73179)/Brie (Addgene #73632 or #73633) libraries from the Broad Institute Genetic Perturbation Platform (GPP) lab. The TKO and Brunello libraries target human coding genes, while the Brie library targets mouse coding genes. The Brunello and Brie libraries are also available in a smaller format targeting the kinome (Addgene #75312–75317). Screens done with these libraries could be sequenced in similar manner by changing primer sequences, but the current protocol is for generating NGS libraries from a GeCKO v2 screen for sequencing on Illumina NextSeq 500.

2 Materials

2.1 Genome-Wide CRISPR-Cas9 Lentiviral Pooled Library

2.1.1 Plasmid Library Amplification

1. GeCKOv2 library plasmid (Addgene pooled library #1000000048 or #1000000049).
2. Endura electrocompetent cells with Recovery Medium (Lucigen Middleton, WI, USA).
3. LB agar bacterial plates.
4. Ampicillin 100 mg/mL: made in sterile water and sterilized by filtration (0.2 μ m filter).
5. Incubator 32 °C.
6. Electroporator (i.e., Gene Pulsar II; Bio-Rad, Hercules, CA, USA).
7. Cell scraper (Corning, Corning, NY, USA).
8. DNA Maxi prep kit.

2.1.2 *Lentiviral Library Packaging in 293FT HEK Cells*

1. Lipofectamine Transfection Reagent (ThermoFisher Scientific, Waltham, MA, USA).
2. PLUS Reagent (ThermoFisher Scientific).
3. pMD2.G plasmid (Addgene #12259).
4. psPAX2 plasmid (Addgene #12260).
5. Opti-MEM (ThermoFisher Scientific).
6. Amicon Ultra-15 Centrifugal Filter Unit with Ultracel-100 membrane (EMD Millipore, Temecula, CA, USA).
7. DMEM/F-12, HEPES (ThermoFisher Scientific).
8. Fetal Bovine Serum.
9. 0.25% Trypsin-EDTA (ThermoFisher Scientific).
10. 293FT human embryonic kidney cells.
11. Humidified incubator, with cells maintained at 37 °C and 5% CO₂.
12. Millex-HP Syringe Filter, 0.45 µm, PES (EMD Millipore).
13. Table top centrifuge.

2.1.3 *Functional Titration of Pooled Lentiviral CRISPR Library*

1. CellTiter-Glo Luminescent Cell Viability Assay (Promega, Fitchburg, WI, USA).
2. Puromycin Dihydrochloride (ThermoFisher Scientific).
3. Luminescence plate reader (i.e., Synergy 2; BioTek, Winooski, VT, USA).
4. Polybrene (EMD Millipore).

2.2 *Genomic DNA Extraction*

2.2.1 *Option A*

1. gDNA Lysis Buffer: 50 mM Tris, 50 mM EDTA, 1% SDS, pH 8. To 75 mL of water add 5 mL of 1 M Tris pH 8.0, 10 mL of 0.5 M EDTA pH 8.0, and 5 mL of 10% SDS.
2. RNase A (100 mg): add 10 mL of gDNA Lysis Buffer to 100 mg of RNase A for 10 mg/mL stock solution. Store at 4 °C.
3. Proteinase K (Qiagen, Hilden, Germany).
4. Ammonium Acetate 7.5 M (Sigma, St. Louis, MO, USA): Dissolve 57.81 g of ammonium acetate in water to a final volume of 100 mL. Sterilize by filtration (0.2 µm filter). Store at 4 °C.
5. Isopropanol.
6. Ethanol.
7. Incubator 55 °C.
8. Table top centrifuge.

2.2.2 Option B

1. QIAamp Blood Midi/Maxi kit (Qiagen) or Quick-gDNA MidiPrep (Zymo Research, Irvine, CA, USA).
2. Microcentrifuge or vacuum manifold.

2.3 PCR NGS Library Prep

1. NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs, Ipswich, MA, USA) or Herculase II Fusion DNA polymerase (Agilent Technologies, Santa Clara, CA, USA) (*see Note 1*).
2. PCR grade water.
3. Thermocycler.
4. PCR1 F primer: AATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTTCG.
5. PCR1 R primer: TCTACTATTCTTTCCCCTGCACTGTTGTGGGCGATGTGCGCTCTG.
6. QIAquick PCR purification kit (Qiagen) or Wizard SV PCR cleanup kit (Promega).
7. PCR2 F :

Primer no.	Illumina P5 and Illumina Seq	Stagger	Inline barcode	Priming site
F01	AATGATACGGCGACCACCGAG ATCTACACTCTTTCCCTACA CGACGCTCTTCCGATCT	t	AAGTAGAG	tcttgtggaaggacgaaacaccg
F02	AATGATACGGCGACCACCGAG ATCTACACTCTTTCCCTACA CGACGCTCTTCCGATCT	at	ACACGATC	tcttgtggaaggacgaaacaccg
F03	AATGATACGGCGACCACCGAG ATCTACACTCTTTCCCTACA CGACGCTCTTCCGATCT	gat	CGCGCGGT	tcttgtggaaggacgaaacaccg
F04	AATGATACGGCGACCACCGAG ATCTACACTCTTTCCCTACA CGACGCTCTTCCGATCT	cgat	CATGATCG	tcttgtggaaggacgaaacaccg
F05	AATGATACGGCGACCACCGAG ATCTACACTCTTTCCCTACAC GACGCTCTTCCGATCT	tcgat	CGTTACCA	tcttgtggaaggacgaaacaccg
F06	AATGATACGGCGACCACCGAGA TCTACACTCTTTCCCTACACG ACGCTCTTCCGATCT	atcgat	TCCTTGGT	tcttgtggaaggacgaaacaccg
F07	AATGATACGGCGACCACCGAGA TCTACACTCTTTCCCTACACG ACGCTCTTCCGATCT	gatcgat	AACGCATT	tcttgtggaaggacgaaacaccg
F08	AATGATACGGCGACCACCGAGA TCTACACTCTTTCCCTACACGA CGCTCTTCCGATCT	cgatcgat	ACAGGTAT	tcttgtggaaggacgaaacaccg

(continued)

Primer no.	Illumina P5 and Illumina Seq	Stagger	Inline barcode	Priming site
F09	AATGATACGGCGACCACCGAGAT CTACTCTTTCCCTACACGAC GCTCTCCGATCT	acgatcgat	AGGTAAGG	tcttgaggaaaggacgaaacaccg
F10	AATGATACGGCGACCACCGAGAT CTACTCTTTCCCTACACGAC GCTCTCCGATCT	t	ACAATGG	tcttgaggaaaggacgaaacaccg

8. PCR2 R primers:

Primer no.	Illumina P7	Index barcode	Illumina seq R	Stagger	Priming site
R01	CAAGCAGAA GACGGCA TACGAGAT	AAGTAGAG	GTGACTGGAGTT CAGACGTGTGC TCTTCCGATCT	t	TCTACTATTCTTT CCCCTGCACTGT
R02	CAAGCAGAA GACGGCAT ACGAGAT	ACACGATC	GTGACTGGAGTT CAGACGTGTGC TCTTCCGATCT	at	TCTACTATTCTTT CCCCTGCACTGT
R03	CAAGCAGAA GACGGCAT ACGAGAT	CGCGCGGT	GTGACTGGAGTT CAGACGTGTG CTCTCCGATCT	gat	TCTACTATTCTTT CCCCTGCACTGT
R04	CAAGCAGAA GACGGCAT ACGAGAT	CATGATCG	GTGACTGGAGTT CAGACGTGTGC TCTTCCGATCT	cgat	TCTACTATTCTTT CCCCTGCACTGT
R05	CAAGCAGAA GACGGCAT ACGAGAT	CGTTACCA	GTGACTGGAGTT CAGACGTGTGC TCTTCCGATCT	tcgat	TCTACTATTCTTT CCCCTGCACTGT
R06	CAAGCAGAA GACGGCAT ACGAGAT	TCCTTGGT	GTGACTGGAGTT CAGACGTGTGC TCTTCCGATCT	atcgat	TCTACTATTCTTT CCCCTGCACTGT
R07	CAAGCAGAA GACGGCAT ACGAGAT	AACGCATT	GTGACTGGAGTT CAGACGTGTGC TCTTCCGATCT	gatcgat	TCTACTATTCTTT CCCCTGCACTGT
R08	CAAGCAGAA GACGGCAT ACGAGAT	ACAGGTAT	GTGACTGGAGTT CAGACGTGTGC TCTTCCGATCT	cgatcgat	TCTACTATTCTTT CCCCTGCACTGT
R09	CAAGCAGAA GACGGCAT ACGAGAT	AGGTAAGG	GTGACTGGAGTTC AGACGTGTGCTC TTCCGATCT	acgatcgat	TCTACTATTCTTT CCCCTGCACTGT
R10	CAAGCAGAA ACGGCATA GAGAT	ACAATGG	GTGACTGGAGT TCAGACGTGTG CTCTTCCGATCT	t	TCTACTATT CTTTCCCC TGCCTGT

3 Methods

The following method is the protocol for generating sequencing libraries for quantification of sgRNA abundance from a genome-wide CRISPR screen using the lentiviral GeCKOv2 library in mammalian cells or tissue on an Illumina NextSeq 500. The general protocol can be adapted for sequencing of custom sgRNA pooled libraries or other ready-made libraries from Addgene.

3.1 Screen Design

Forward (or functional) genomic screens can fundamentally be carried out in two ways. One is dropout (or negative selection) screening in which a gene knockout results in a selective disadvantage in that cell such as decreased proliferation in cancer cells. The other is enrichment (or positive selection) screening in which gene knockout results in selective advantage for cells such as drug or toxin resistance.

After the appropriate phenotype has been selected and a dropout or resistance screen has been designed, two other critical factors to take into account in the design of genome-wide screens is (1) the depth of coverage of each sgRNA guide and (2) the number of sgRNA guides per gene. Experience from previous genome-wide shRNA screens suggested that dropout screens require much higher coverage than resistance screens with the recommendation for at least 500–1000× coverage of each individual element for dropout screens to obtain enough signal over noise [15]. Because of the different mechanisms of CRISPR-Cas9 resulting in gene knockouts rather than knockdown, it seems that signal-to-noise ratio is higher than in shRNA screens and the initial CRISPR-Cas9 genomic dropout screens have been carried out at a minimum of 200–300× coverage [9, 11]. For shRNA screens, because of the significant off-target effects of shRNAs, one typically needed to include many shRNAs per gene (typically 5–25 shRNAs per gene) to overcome the noise generated by off-target effects. CRISPR screening appears to have less off-target effects, and the reason to include multiple sgRNAs per gene for CRISPR screening appears to be the variability in the efficiency of each individual sgRNA rather than the off-target noise such that increasing the amount of sgRNAs appears to increase the sensitivity of the screen rather than the specificity [10, 16]. One of the largest hurdles in performing genome-wide screening is providing for adequate coverage of the library by using appropriate amounts of starting cells and maintaining that coverage throughout the experiment. For example, if using the entire human GeCKOv2 library containing 122,411 sgRNA guides, one would need to transduce 2×10^8 cells at 30% efficiency and maintain 6×10^7 cells at each passage and harvesting time point for each biological replicate. This amount of cells could be prohibitive in certain situations. An alternative emerging

strategy is to decrease the amount of sgRNAs performed at the genome-wide level (to 3–4 sgRNAs) in the primary screen and to use relaxed cutoffs for hit selection and perform smaller targeted secondary screens using additional sgRNAs per gene [16].

3.2 Genome-Wide CRISPR-Cas9 Lentiviral Pooled Library

3.2.1 Plasmid Library Amplification

3.2.2 Lentiviral Library Packaging in 293FT HEK Cells

1. Obtain GeCKOv2 library from Addgene delivered as half-libraries (A and B each with three sgRNAs per gene) in 20 μL at a concentration of 50 ng/ μL .
2. Follow provided protocol from Addgene and Zhang lab for amplification of library using pooled electroporation. (https://www.addgene.org/static/cms/filer_public/b5/fd/b5fde702-d02c-4973-806f-24ac28b2a15a/geckov20_library_amplification_protocol_1.pdf).
1. Culture 293FT cells in DMEM with 10% FBS (D10 media) without antibiotics, prepare a 10 cm plate about 90% confluent day of transfection.
2. Wash the cells with 3 mL room temperature PBS.
3. Aspirate PBS, add 3 mL Trypsin/EDTA and incubate at 37 °C for 2 min.
4. Add 3 mL D10 to inactivate Trypsin/EDTA and collect the cells by centrifugation (1000 rpm for 10 min).
5. Resuspend the cells in 5 mL pre-warmed DMEM media without serum.
6. Plate the cells in a 10 cm tissue culture plate and place in an incubator.
7. Prepare plasmid/PLUS mixture by adding 900 μL room temperature Opti-MEM in a sterile 1.5 mL microcentrifuge tube. Add 12 μg of amplified GeCKO library plasmid, 9 μg psPAX2 plasmid, 6 μg pMD2.G plasmid, and 48 μL of PLUS reagent.
8. Prepare lipofectamine mixture by adding 60 μL of lipofectamine reagent to 900 μL of room temperature Opti-MEM in a sterile 1.5 mL microcentrifuge tube.
9. Vortex mixtures briefly and incubate at room temperature for 5 min.
10. Combine plasmid/PLUS mixture with lipofectamine mixture. Vortex briefly and incubate at room temperature for 15 min.
11. Add plasmid/lipofectamine mixture dropwise to cell suspension from **step 6** and place in an incubator for 4–6 h.
12. Aspirate media and replace with 10 mL fresh D10 media.
13. Harvest the supernatant after 60 h into a 50 mL conical tube.
14. Remove cellular debris by centrifugation at 1600 $\times g$ at 4 °C for 10 min.
15. Filter the supernatant through a 0.45 μm syringe filter.

16. Concentrate lentivirus using Amicon Ultra-15 centrifugal filter by adding filtered viral supernatant into reservoir and centrifuging at $2850 \times g$ for 40 min at 4 °C.
17. Aliquot virus and store frozen at -80 °C.

*3.2.3 Functional Titration
of Pooled Lentiviral CRISPR
Library*

1. Harvest target cells and count cells and resuspend cells at a concentration of $2-3 \times 10^6$ cells/mL in appropriate media for the cell type.
2. Add 1 mL cells into each well of a 12-well tissue culture plate and add polybrene to a final concentration of 8 $\mu\text{g}/\text{mL}$ per well.
3. Add different amounts of concentrated virus (usually in the range of 0.5–20 μL for virus generated from GeCKOv2 1 plasmid system, titers are generally higher if using the 2 plasmid system) to each well.
4. Spinfect by centrifugation at $700 \times g$ for 1–2 h at 32 °C.
5. Remove media by aspiration and replace with fresh media without polybrene.
6. Incubate for 6 h or overnight then collect cells by washing gently with 500 μL PBS per well.
7. Aspirate PBS and add 500 μL Trypsin/EDTA per well and incubate until cells detach. Inactivate by adding 1 mL appropriate media for cell type with serum per well.
8. Centrifuge the cells at 1000 rpm for 5 min and aspirate media.
9. Resuspend in cells collected from each well in 10 mL of appropriate media and add 1 mL of cell suspension to duplicate wells in a 6-well tissue culture plate. Add another 3 mL of appropriate media to each well.
10. In each duplicate, add puromycin to one replicate well at a concentration that results in no surviving cells after 3 days.
11. After 3 days of puromycin selection count viable cells using CellTiterGlo by aspirating media and adding CellTiterGlo (diluted 1:1 in PBS) to each well.
12. Cover the plates with foil and shake for 2 min followed by incubation for 10 min at room temperature.
13. Read luminescence on plate reader and compare with control untreated plate.
14. Choose the concentration of virus resulting in 20–40% of cell survival compared to untreated replicate well (this corresponds to a MOI of around 0.3–0.4).
15. Scale up to necessary amount of cells to conduct primary screen by increasing the number of wells spinfected with $2-3 \times 10^6$ cells/well.

3.3 Harvesting Genomic DNA

1. For each time point and replicate in the screen, collect and freeze the amount of cells required to maintain predetermined coverage (6×10^7 cells for $500\times$ coverage of full GeCKOv2 human library) at -20°C in 15 mL conical tubes with up to 200 mg tissue or 3×10^7 cells per tube.
2. Thaw cells until the pellet can be dislodged easily by flicking tube, proceed to genomic DNA extraction using salt precipitation (*option A*) or commercial silica-membrane kits (*option B*).

3.3.1 Option A

1. The salt precipitation method used in the screen conducted by the Zhang lab [17] provides high yield genomic DNA with consistent results in sequencing. This option is more cost-effective and simpler when extracting DNA from large amounts of cells or tissues compared to using the commercial silica membrane kits.
2. For 100–200 mg of tissue or 3×10^7 – 5×10^7 cells in a 15 mL conical tube, add 6 mL gDNA lysis buffer and 30 μL of proteinase K to tissue/cells.
3. Incubate at 55°C overnight.
4. After complete lysing of cells/tissue, add 30 μL of 10 mg/mL RNase A.
5. Incubate at 37°C for 30 min, then cool on ice.
6. Add 2 mL of cold 7.5 M ammonium acetate and vortex for 20 s.
7. Centrifuge $>4000 \times g$ for 10 min.
8. Transfer the supernatant to a new tube and add 6 mL 100% isopropanol. Mix by inverting tube 50 times.
9. Centrifuge $>4000 \times g$ for 10 min. Discard the supernatant, add 6 mL of 70% ethanol, and mix by inverting tube ten times.
10. Centrifuge $>4000 \times g$ for 5 min, discard the supernatant by decanting and aspirating with P200 tip.
11. Air dry 10–30 min until the pellet becomes slightly translucent.
12. Resuspend in 500 μL of TE.
13. Incubate at 65°C for 1 h, then at room temperature overnight.
14. Measure gDNA concentration on nanodrop 2000 next day.

3.3.2 Option B

1. Extract gDNA from frozen cell pellets using QIAamp Blood midi kit according to the manufacturer's protocol.
2. Elute gDNA in EB or TE buffer and incubate at room temperature overnight.
3. Measure gDNA concentration on nanodrop 2000 next day.

3.4 PCR NGS Library Prep

1. NGS libraries are generated by a two-step PCR where the first PCR amplifies the sgRNA region utilizing primers recognizing constant lentiviral integration sequence and a second

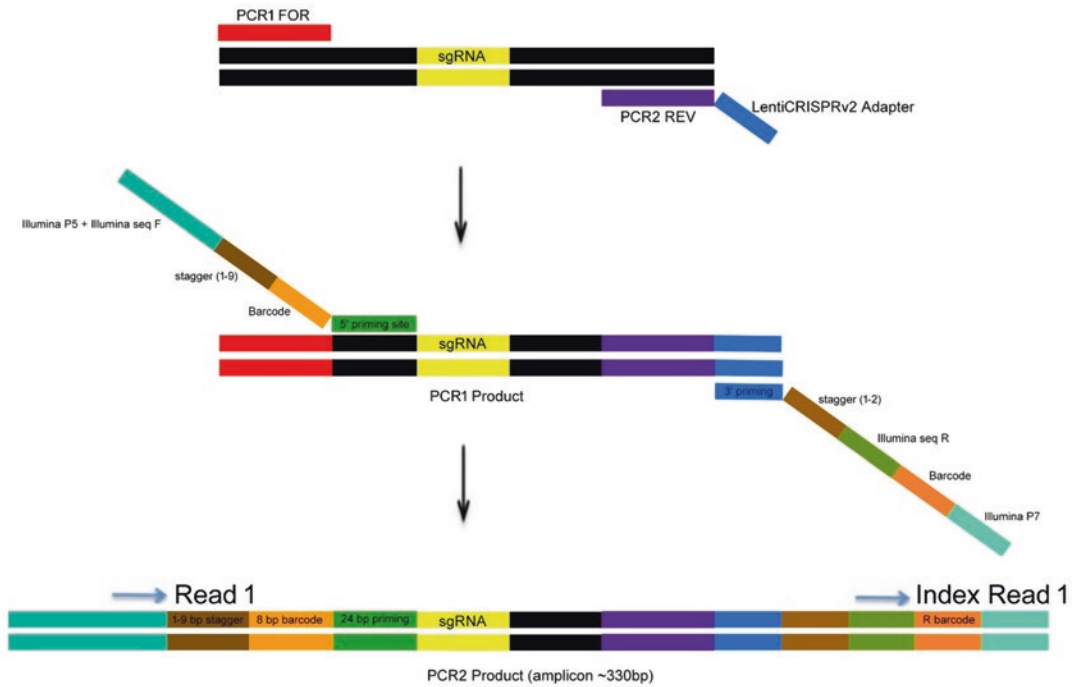


Fig. 1 Schematic representation of two step PCR for the generation of NGS libraries (Subheading 3.4)

PCR adds illumina i5 and i7 sequences as well as barcodes for multiplexing directly in front of the variable 20 bp sgRNA sequences (Fig. 1). sgRNA libraries generated with these primers can then be sequenced on Illumina NextSeq (1 × 75) single indexed read.

3.4.1 PCR1

1. It is important to maintain library coverage during the library preparation. For the generation of libraries for the amplified plasmid, 10 ng of starting material is more than adequate. For libraries generated from harvested genomic DNA, using the estimation of 7 μg of gDNA per 10⁶ cells, an adequate number of PCR1 reactions will need to be performed. Using Herculase II Fusion DNA Polymerase, we can generally use up to 10 μg of gDNA per 100 μL PCR reaction. Using NEBNext Q5 Hot start HiFi Polymerase, we can generally use up to 5 μg of gDNA per 100 μL PCR reaction. So for maintaining 500× coverage of a full GeCKOv2 library, one would need to perform 40 separate 100 μL PCR1 reactions with Herculase II Fusion DNA Polymerase and 85 separate 100 μL PCR1 reactions with NEBNext Q5 Hot Start HiFi Polymerase.
2. PCR1 primers can be ordered from IDT at the smallest scale (25 nM, standard desalted) and resuspended in TE for a final concentration of 10 μM.

3. Prepare 100 μL PCR reactions as follows: (a) For Herculase II Fusion DNA Polymerase: prepare reactions on ice: 20 μL Herculase 5 \times Buffer; 1 μL dNTP (100 mM, 25 mM each dNTP); 2.5 μL PCR1 F primer (10 μM); 2.5 μL PCR1 R primer (10 μM); 1 μL Herculase II Fusion Enzyme; 10 ng plasmid or up to 10 μg genomic DNA; PCR grade water to 100 μL total volume. (b) For NEBNext Q5 HotStart HiFi PCR Mastermix: 50 μL NEBNext Q5 HotStart HiFi PCR mastermix; 1.5 μL PCR1 F primer (10 μM); 1.5 μL PCR1 R primer (10 μM); 10 ng plasmid or up to 5 μg genomic DNA; PCR grade water to 100 μL total volume.
4. Place the tubes in a thermocycler and run the following programs:

For Herculase II Fusion Enzyme:

Cycles	Temp.	Duration
1	95 $^{\circ}\text{C}$	2 min
18–22 (see Note 2)	95 $^{\circ}\text{C}$	10 s
	60 $^{\circ}\text{C}$	20 s
	72 $^{\circ}\text{C}$	35 s
1	72 $^{\circ}\text{C}$	3 min

For NEBNext Q5 Enzyme:

Cycles	Temp.	Duration
1	98 $^{\circ}\text{C}$	2 min
12–18 (see Note 2)	98 $^{\circ}\text{C}$	10 s
	60 $^{\circ}\text{C}$	30 s
	65 $^{\circ}\text{C}$	45 s
1	65 $^{\circ}\text{C}$	5 min

5. Purify PCR products using QIAquick PCR purification or Wizard SV PCR cleanup. Resuspend each PCR reaction in 100 μL of TE. Pool PCR1 reactions together for each experimental condition and use 10 μL of PCR1 product as a template for PCR2.

3.4.2 PCR2

1. To maintain library complexity, we will need to perform one PCR2 reaction per 10^4 constructs in library (so 13 reactions for GeCKOv2 A and B library for each experimental condition) with technical duplicates.
2. PCR2 primers are longer and more costly than PCR1 primers and can be synthesized by IDT as Ultramer DNA Oligos at 4nmole scale and resuspended in TE to a final concentration of 10 μM .

3. Assemble 50 μL PCR2 reactions using 10 μL of purified pooled PCR1 product as template. Each experimental condition can be barcoded with unique F and R primers for multiplexing on the same NextSeq flowcell.

For Herculase II Fusion DNA Polymerase (prepare reactions on ice): 10 μL Herculase 5 \times Buffer; 0.5 μL dNTP (100 mM, 25 mM each dNTP); 1.25 μL PCR2 F primer (10 μM) (F01–F10); 1.25 μL PCR2 R primer (10 μM) (R01–R10); 0.5 μL Herculase II Fusion Enzyme; 10 μL PCR1 product; PCR grade water to 50 μL total volume.

For NEBNext Q5 HotStart HiFi PCR Mastermix: 25 μL NEBNext Q5 HotStart HiFi PCR mastermix; 1.5 μL PCR2 F primer (10 μM) (F01–F10); 1.5 μL PCR2 R primer (10 μM) (R01–R10); 10 μL PCR1 product; PCR grade water to 50 μL total volume.

4. Place the tubes in a thermocycler and run the following programs:
For Herculase II Fusion Enzyme:

Cycles	Temp.	Duration
1	95 °C	2 min
18–22 (<i>see Note 2</i>)	95 °C	10 s
	60 °C	20 s
	72 °C	35 s
1	72 °C	3 min

For NEBNext Q5 Enzyme:

Cycles	Temp.	Duration
1	98 °C	2 min
12–18 (<i>see Note 2</i>)	98 °C	10 s
	60 °C	30 s
	65 °C	45 s
1	65 °C	5 min

5. Purify PCR products using QIAquick PCR purification or Wizard SV PCR cleanup. Resuspend each PCR reaction in 50 μL of TE and pool products with the same barcode but keeping technical replicates separate.
6. Run PCR product on 2% E-gel and gel purify approx. 370 bp band.
7. Quantify by Qubit or bioanalyzer and pool barcoded samples for sequencing on Illumina NextSeq 500.

4 Notes

1. It is important to use a high-fidelity, GC unbiased polymerase to obtain accurate NGS libraries for quantification. In addition to Herculanase II Fusion and Q5, many labs have had success with Kapa HiFi enzyme. Due to its GC bias, we would avoid using Phusion polymerases. We have found Herculanase II Fusion is able to tolerate the largest amount of genomic DNA. In order to help with the large amounts of genomic DNA, one can also supplement additional $MgCl_2$ up to a final concentration of 4 mM.
2. Some optimization of cycle numbers is required for PCR1 and PCR2. Given the quantitative goal of NGS library generation, keeping the PCR reactions in the linear range is desired. For Herculanase II Fusion, we have found that around 18–20 cycles of PCR1 and 18–20 cycles of PCR2 is a good place to start. Excessive cycle numbers of PCR2 results in a large MW smear seen on agarose gels. Using Q5 enzyme, even lower cycle numbers can be used with around 15–18 cycles for PCR1 and 12–15 cycles for PCR2.

Acknowledgments

We thank the Rana lab members, John Shimishata, and Steven Head at The Scripps Research Institute Genomic Core. This work was supported in part by grants from the National Institutes of Health to T.M.R., E.Y. is supported by the National Cancer Institute of the National Institutes of Health under award number T32CA121938. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

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* 337:816–821
2. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM (2013) RNA-guided human genome engineering via Cas9. *Science* 339:823–826
3. 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* 339:819–823
4. Doudna JA, Charpentier E (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* 346:1258096
5. Hsu PD, Lander ES, Zhang F (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell* 157:1262–1278
6. Sander JD, Joung JK (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol* 4:347–355
7. Sternberg SH, Doudna JA (2015) Expanding the biologist's toolkit with CRISPR-Cas0. *Mol Cell* 58:568–574

8. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM (2015) Identification and characterization of essential genes in the human genome. *Science* 350:1096–1101
9. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S, Mero P, Dirks P, Sidhu S, Roth FP, Rissland OS, Durocher D, Angers S, Moffat J (2015) High-resolution CRISPR screen reveal fitness genes and genotype-specific cancer liabilities. *Cell* 163:1515–1526
10. Morgens DW, Deans RM, Li A, Bassik MC (2016) Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes. *Nat Biotechnol* 34:634–636
11. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* 343:84–87
12. Wang T, Wei JJ, Sabatini DM, Lander ES (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science* 343:80–84
13. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Gulmaras C, Panning B, Pioegh HL, Bassik MC, Qi LS, Kampmann M, Weissman JS (2014) Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* 159:647–661
14. Virreira Winter S, Zychlinsky A, Bardeol BW (2016) Genome-wide CRISPR screen reveals novel host factors required for *Staphylococcus aureus* α -hemolysin-mediated toxicity. *Sci Rep* 6:24242
15. Kampmann M, Bassik MC, Weissman JS (2014) Functional genomics platform for pooled screening and mammalian genetic interaction maps. *Nat Protoc* 9:1825–1847
16. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, Virgin HW, Listgarten J, Root DE (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* 34:184–191
17. Chen S, Sanjana NE, Zheng K, Shalem O, Lee K, Shi X, Scott DA, Song J, Pan JQ, Weissleder R, Lee H, Zhang F, Sharp PA (2015) Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell* 160:1246–1260

Gene Profiling and T Cell Receptor Sequencing from Antigen-Specific CD4 T Cells

Marie Holt, Anne Costanzo, Louis Gioia, Brian Abe, Andrew I. Su, and Luc Teyton

Abstract

The paucity of pathogenic T cells in circulating blood limits the information delivered by bulk analysis. Toward diagnosis and monitoring of treatments of autoimmune diseases, we have devised single-cell analysis approaches capable of identifying and characterizing rare circulating CD4 T cells.

Key words CD4 T cells, Single-cell analysis, TCR sequences, Autoimmunity, Peripheral blood, Microfluidics

1 Introduction

One of the hallmarks of the immune system is the heterogeneity of each of the cell populations it is composed of. The classical way of addressing this heterogeneity has been to subset each population with a series of cell surface markers amenable to fluorescence-activated cell sorting techniques (FACS). Counterintuitively, FACS analysis, a technique based on single-cell isolation principles, is usually followed by population analysis and the ignorance of heterogeneity within the studied population. Like many of our colleagues, we have been questioning what heterogeneity might mean in physiology and diseases. When studying T cells for instance, we have known for decades that plasticity allows interconversion of effector functions when environmental cues change. The extent of this variability and the specificity of each effector program in the context of antigen-driven immune responses must be studied to devise better strategies for vaccination and better understanding of autoimmunity. Attempts at studying gene expression or T cell receptor (TCR) usage of single T cells are not new but have met very limited success until recently [1, 2]. The development of microfluidic technologies and better molecular biology reagents

have revolutionized the field over the past 5 years. We have moved from days when the efficiency of paired sequencing of α and β TCR chains was exceptional [3], to a situation where this technique is routine and whole exome sequencing from single cells is the new frontier [4]. Like for FACS analysis, a small number of platforms are now available for the genomic interrogation of single cells. Droplet-based microfluidic is one system [5], whereas purely microfluidic system is the other [6]. We used the latter, mainly because it entered the field first and we committed to this approach early on. In any case, we have used this approach to study antigen-specific CD4⁺ T cells in the context of vaccines and autoimmunity. For efficiency, and in order to collect the most information from a single cell, we have developed a workflow that allows the profiling of 96 genes by quantitative PCR (qPCR), as well as the sequencing of paired α - β TCR chains from the same original reverse transcription of pMHC-FACS sorted T cells. The technique has proven robustness, high reproducibility, and can be used for routine analysis of antigen-specific T cells [7]. For our purpose, we have been using a Fluidigm Biomark HD system and its two main components, the Dynamic Array for qPCR, and the Access Array for the making of libraries for next-generation sequencing.

2 Materials

2.1 Cell Sorting Components

1. Allegra X-14R centrifuge (Beckman Coulter, Inc., Brea, CA, USA).
2. PBS: Dulbecco's Phosphate-Buffered Saline, 1 \times (Corning, Corning, NY, USA).
3. Cell strainer, 70 μ m (ThermoFisher Scientific, Waltham, MA, USA).
4. FACS buffer: 2% FCS, 2 mM EDTA in PBS, sterile filtered.
5. RBC Lysis Buffer: 0.165 M NH₄Cl in deionized water, sterile filtered.
6. Fc block: 20 μ g/mL in FACS buffer.
7. Tetramers: In a molar ratio of 1:5, biotinylated MHC molecules expressed with peptides of interest are mixed in sterile PBS with PE-labeled streptavidin (Life Technologies, Carlsbad, CA, USA) (*see Note 1*). The reaction is incubated overnight, in darkness, at room temperature.
8. Antibodies: For the samples the following antibodies are mixed together 1:200 in FACS buffer: FITC anti-mouse CD3 ϵ , APC/Cy7 anti-mouse CD4, APC anti-mouse/human CD11b, APC anti-mouse/human CD45R/B220, APC anti-mouse CD49b, and APC anti-mouse CD8a (all from BioLegend, San Diego, CA, USA). For the compensations the following antibodies are

mixed individually 1:200 in FACS buffer: FITC anti-mouse CD3 ϵ , APC/Cy7 anti-mouse CD4, APC anti-mouse/human CD45R/B220, and PE anti-mouse CD4 (BioLegend, San Diego, CA, USA).

9. Hard-Shell 96-Well Semi-Skirted PCR plates, High-Profile (BioRad, Hercules, CA, USA).
10. RT Mix Solution 1 (master mix for one plate): 144 μ L 5 \times VILO reaction mix (SuperScript VILO cDNA Synthesis Kit, ThermoFisher Scientific), 36 μ L 20 U/ μ L SUPERase-In (ThermoFisher Scientific), 30 μ L 10% NP40 (ThermoFisher Scientific), and 390 μ L Nuclease-free water (TEKnova, Hollister, CA, USA).
11. FACSARIA I cell sorter (BD Biosciences, San Jose, CA, USA).

2.2 Reverse Transcription and Pre-amplification Components

1. T100 Thermal Cycler (BioRad).
2. RT Mix Solution 2 (master mix for one plate): 18.0 μ L 10 \times SuperScript Enzyme Mix (SuperScript VILO cDNA Synthesis Kit, ThermoFisher Scientific), 14.4 μ L T4 Gene 32 Protein (New England Biolabs, Ipswich, MA, USA), and 87.6 μ L Nuclease-free water (TEKnova).
3. 10 \times STA (Specific Target preAmplification) Primer Mix: pool all primer pairs (100 μ M) (*see* **Notes 2** and **3**) in equal volumes, and dilute in 1 \times DNA Suspension Buffer (TEKnova) giving a final concentration of 500 nM of each primer. Make several aliquots for one time use, and store at -20°C .
4. STA Reaction Mix (master mix for one plate): 780 μ L TaqMan PreAmp MasterMix (Life Technologies), 156 μ L 10 \times STA Primer Mix, and 7.8 μ L 0.5 M EDTA, pH 8.0 (Life Technologies).
5. Exonuclease (master mix for one plate): 72 μ L Exonuclease I Reaction Buffer (10 \times), 144 μ L Exonuclease I (20 U/ μ L) (New England Biolabs), and 504 μ L Nuclease-free water (TEKnova).

2.3 qPCR Components

1. IFC Controller HX (Fluidigm, San Francisco, CA, USA).
2. 96.96 Dynamic Array IFC (Fluidigm).
3. Control line fluid (150 μ L) (Fluidigm).
4. Sample Pre-Mix Solution (master mix for one plate): 480 μ L 2 \times Sso Fast EvaGreen Supermix With Low ROX (BioRad) and 48 μ L 20 \times DNA Binding Dye Sample Loading Reagent (Fluidigm).
5. qPCR Primer Plate: 96 primer pairs (100 μ M) (DELTAgene Assays, Fluidigm (*see* **Table 1**)) are separated in a 96-well master plate. Aliquot into working plates for use up to five times and store at -20°C (*see* **Note 4**).

6. Assay Mix Solution (master mix for one plate): 540.0 μL 2 \times Assay Loading Reagent (Fluidigm) and 486.0 μL 1 \times DNA Suspension Buffer (TEKnova).
7. BioMark HD instrument (Fluidigm).

2.4 Access Array Components

1. IFC controller AX (Fluidigm) (*see Note 5*).
2. 48.48 Access Array IFC (Fluidigm).
3. Control line fluid (300 μL) (Fluidigm).
4. 1 \times Access Array Harvest solution (Fluidigm).
5. 1 \times Access Array Hydration Reagent v2 (Fluidigm).
6. Internal TCR 5' Primer Plate (20 \times): In half of a 96-well plate add to each well 5 μL 20 \times Access Array Loading Reagent (Fluidigm) and 87 μL Nuclease-free water (TEKnova). Add 8 μL of each TCR specific internal 5' primer (50 μM) (IDT, Coralville, IA, USA) (*see Table 3*) to the individual wells, giving a final concentration of each primer of 4 μM (*see Notes 3 and 6*). Aliquot into working plates for use to up to five times and store at -20°C .
7. Sample Pre-Mix Solution (master mix for one sample plate): 60 μL 10 \times FastStart High Fidelity Reaction Buffer Without MgCl_2 (Roche, Basel, Switzerland), 108 μL 25 mM MgCl_2 (Roche), 30 μL DMSO (Roche), 12 μL 10 mM PCR Grade Nucleotide Mix (Roche), 6 μL 5 U/ μL FastStart High Fidelity Enzyme Blend (Roche), 30 μL 20 \times Access Array Loading Reagent (Fluidigm), 186 μL Nuclease-free water (TEKnova).
8. Barcoded Internal TCR 3' Primer Plate: In half of a 96-well plate make 48 pairs of 3' primers with matching barcodes (IDT) binding the constant regions of TCR α and TCR β (TRAC and TRBC) respectively (*see Table 4 and Note 3*). Combine equal volumes of each primer (100 μM) giving a final concentration of 50 μM . Aliquot into working plates for use to up to five times and store at -20°C .

2.5 Library Construction Components for Illumina Sequencing

1. AMPure XP magnetic beads (Beckman Coulter, Carlsbad, CA, USA).
2. Magnetic Separator.
3. 2100 Bioanalyzer instrument and High-Sensitivity DNA chip (Agilent, Santa Clara, CA, USA).
4. 80% ethanol.
5. Nuclease-Free water.
6. Qubit $-$ Fluorometer and ds DNA high sensitivity assay (ThermoFisher Scientific, Waltham, MA, USA) (similar fluorescent DNA quantitative detection methods can be substituted).
7. NEBNext Ultra DNA Library Prep Kit for Illumina.

Table 1
96 primer pairs for pre-amplification and qPCR

Target	Forward	Reverse	Target	Forward	Reverse
Bel6	GGGAAACCCAGTCAG AGTA	CTCAGAGAAACGGCAGT CAC	Aim2	CTGCTACAGAACTCTGTCTCA	TAGCTTTCAGCACC GTGACA
Ccr2	TGAGGCTCATCTTTGCC ATCA	GGATTCTGGAAGGT GGT CAA	Bel2	ATGTGTGTGGAGAGCGTCAA	GATGCCGGTTCAGGT ACTCA
Ccr3	CTGGACTCATAAAGG ACTTAGCA	GTGGTGCCCACTCATAT TCA	Ccr1	TCCTCAAAGGCCCCAGAAACA	GCTGAGGAACTGGT CAGGAA
Ccr4	GACTGTCCCTCAGGATC ACTTTCA	CCTGGGTGGTGTCTGT GAC	Ccr7	GTGGTGGCTCTCCTTTGTCA	GGTATTCTCGC CGATGTAGTCA
Ccr5	GGAGGTGAGACATCC GTTCC	GGTCGGAACTGACCCCTT GAAA	Pdl-1	CAGCCCTGCTGTCACTTTGCTA	GACGTTGCTGCCATA CTCCA
Ccr6	AAGGCACATATGCGG TCAAC	CCTGGACGATGGCAATG TAC	Cd44	TTCTTCGATGGACCCGGTTA	TACTCGCCCTTCTT GCTGTA
Cd28	CTGCTGTTCTTTGGCTC TCAAC	GGGCGACTGCTTTTACCA AAA	Ceacam1	GCGACTGTGCGATTTCATGTA	AGGTCAGGGTCACA GAGTCTA
Cd3e	TGCTACACACCAGCC TCAAA	AGGTCCACCTCCACACA GTA	Cxcl10	GGGCCATAGGGAAGCTTGAA	GGATTCAGACATCT CTGCTC ATCA
Cd4	AAGGACACTGCATC AGGAA	CCCATCACCTCACAGGTC AA	Cxcr3	ACCAGCCAAAGCCATGTACC	GGGAGAGGTGCTG TTTTCCA
Cd40	CTATGGGGCTGCTTG TTGAC	TCGTGGAGGTACTGTTT GTCA	Cxcr4	GGTAACCAACCACGGCTGTA	CAGGGTTCCTTGTG GAGTCA

(continued)

Table 1
(continued)

Target	Forward	Reverse	Target	Forward	Reverse
Cd80	AGTCGTCGTCATCGTT GTCA	GTTTGTTCCTCTGCT TGCCCTCA	Fyn	TGGCTGGGTTGATTGAAG ACA	GGGCTGTCCACTTA ATGGGAA
Cd86	CATGGGCTTGGCAAT CCTTA	CATTGAAATAAGCTTGGGT CTCC	Icos	TGACCCACCTCCTTTTCAA GAA	TACGGGTAGCCAGA GCTTCA
Cd8a	CAGCAAGGAAAACGAAG CTAC	GCAGCACTGGCTTGGTA GTA	Ifi44	TCITGGTGGGCTGTGAT GAA	TCATCCTTGGCCT TGATGGAA
Ctla4	GGACTTGGCCCTTTTGT AGCC	CTGAAGGTGGGTCCACT GTA	Ifi441	GCAAACATGACAGAAAATG TGAC	CAACCTTCGCTCT GAAAGCATAA
Foxp3	CCCACACCTCTTCTTCC TTGAA	GACGGTGCCACCATGA CTA	Ifit1	GCTACCACTTTTACAGC AACC	AGTGACATCTCAGC TGAAGCA
Gapdh	CAAGGTCATCCCAGAG CTGAA	CAGATCCACGACGGAC ACA	Ifit3	TTTTCTGGCACCATGA ACC	TCCACAGCACATC TGTCTCA
Gata4	GTAATGCCTGCGGCC TCTA	TGGTTTGAATCCCCTCC TTCC	Ifngr1	CTGGGAATACCAGAACATG TCAC	TGCAGGAATCAG TCCAGGAA
Gsk3a	GAACTGGTGGCCATC AAGAA	ATTGCAGTGGTCCAGCT TAC	Il12rb	GCGTTGAGAAGACATCGT TCC	TGGAAAACCCTGT AGCAACTCA
Gsk3b	GCAGCCTTCAGCTTT TGGTA	GGAGTTGCCACTACTGTG GTTA	Il18r1	AAGAGGACAGCTCAGACC CTAA	GAAGCATGCAGT TTGCCTTCA
Hprt	CAGTACAGCCCCAAA ATGGTTA	AGTCTGGCCTGTATCCA ACA	Il27r	GGTCCCAACCTTTTCACTT CAC	AAACCCACACAGG GACAGAAA

Icam1	AGGGCTGGCATTGTT CTCTA	TGTCGAGCTTTGGGAT GGTA	Irf1	TACCTGGGTCAGGACTTG GATA	TCAGAGAGACT GCTGCTGAC
Ifng	CCACGGCACAGTCA TTGAAA	GCCAGTTCCCTCCAGATAT CCAA	Irf2	GTGGCTGGAGGAGCAGAT AAA	GCATCCAGGGG ATCTGGAAA
Il10	AAAGGACCAGCTGGAC AACA	TAAGGCTTGGCAACCCAA GTA	Irf4	TCCCCATTGAGCCAAGC ATA	CGAGGATGTCC GGTAATACA
Il12b	ATCGTTTTGCTGGTG TCTCC	GGAGTCCAGTCCACCT CTAC	Irf7	GATCCGCATAAAGGTGTAC GAAC	TGCTGAGGCTCA CTTCTTCC
Il17A	CAGACTACCTCAACCG TTCCA	CACTGAGCTTCCCAGATC ACA	Isg15	GGACGGTCTTACCCTTT CCA	TCGCTGCAGTTCT GTACCA
Il1r2	AGTGCAGCAAGACTC TGGTA	AGACCTTGAGTTCCACAG ACA	Jak1	TGGCCCGTTTCATCAAGC TTA	CAGGGGATTCGCT CTATGCA
Il2	CCCAGGATGCTCACCTT CAAA	CCGCAGAGGTCCAAGT TCA	Jak2	TGCCCGCAGGACAAAAGAAT Acb	TGCTCTCCGTCAA GGATTCA
Il-21	GATCCTGAACCTTCTAT CAGCTCCA	GGCCTTCTGAAAAACAGG CAAA	Ly6e	GTTTCATGCCAGGAGAAA GACC	AGGGTGTAGCCAA GGTTGAC
Il25	CTCTCTCAGAAGGCC TGTC	CCCACGATCATTTGCCA AGAA	Map2k6	ATGCCGGTTGCCAAACCA TAC	TGTCAGACTTCAC ACTGTA CCC
Il27	GGCCAGGTGACAGGAGAC	CAGGAAACAGCTTGTAC CAGAA	Mapk8	GGGAGAAAATGGTTTGCCA CAA	AGGTGTTCCGAGC TGTCAA

(continued)

Table 1
(continued)

Target	Forward	Reverse	Target	Forward	Reverse
II2ra	TGCGTTGCTTAGGAA ACTCC	CTGGTGTTCAAAGTTGAGC TGTA	Mx1	GAGATGACCCAGCACCT GAA	GGATAATCAGAGG GATCTG TCTCC
II3	TCGTGGAAGCCAAG GAGAA	AGATGTAGGCCAGGCAAC AGTTA	Nur77	CAATGCTTTCGTGTGTCAGC ACTA	TGTTTGCCAGGCA GATGTAC
II4	ACGGAGATGGATGT GCCAAA	GAAGCACCTTGGGAAGCC CTA	Oas1b	GTGCTGCCAGCCTATGA TTTTA	CGATAACTTGCCCT CCTTCC
II4ra	AACATCTCCAGAGAGGA CAACC	CTCAGCCCTGGGTTCCCTT GTA	Oas2	CTGTACTCTCCCAGCCT GAA	GTGGCTTGGAGTG ACGAAAA
II5	GATGAGGCTTCCCTGT CCCTA	TTCAGTATGTCTAGCCCC TGAA	Oas1l	GTCATCGAGGCCTGTG TCA	TCTGTGGGTCCA GGATGATA
II5ra	CTGCTGAACTCAAAG CTCCA	AGTGGGTGTGGCTACTT ACA	Pd1	CTGGAAGCAAGGACGA CAC	CTGGAAGTCCAGC TCCTCATA
II6	CGATGATGCACTTGC AGAAA	ACTCCAGAAAGACCAGAG GAA	Rsad2	GAGCAATGGCAGCCCTTA TCC	TGTCGCAGGAGAT AGCAAAGAA
II7	GCTGCAGTCCCAGT CATCA	AGGCAGCAGAACAAGGA TCA	Socs3	AAGCCGGAGATTTCG CTTC	TGGCTGCAGCTG CTTCG
II7r	AAAGCATGATGTGGC CTACC	GGGATTGTGTTCTTTGTG TGGAA	Stat1	GCAGGTGTTGTGATC GAAC	ATGCACGGCTGT CGTTCCTA
Nfkb1	ACCGTATGAGCCTGT GTTCA	GTAGCCCTGTGTCTTCT GTCA	Stat3	TGGGCATCAATCCTGTG GTA	CCAATTGGGGC TTAGTGAA
Ppara	AGGGTACCACTACGG AGTTCA	ACACCAGCTTCAGCCGA ATA	Stat4	CCCAAGGAGATGAAGTG CAGTA	ATGGAATGCA ACTCCTCTGTCA

Pparg	ACCCAAATGGTTGCTG ATTACA	AGGTGGAGATGCAGGT TCTA	Stat5	AGCCAGGACCACAATG CTA	CCTTGT CAGGCAC AGCAAA
Pparg1a	AAACCACACCCACA GGATCA	GCTCTTCGCTTTATTGCT CCA	Tgfb2	TCTGTGAGAA GCCGCAT GAA	GGCAAA CCGT CTCCAGAGTAA
Pten	GAGACATTATGACACC GCCAAA	AAGTCTAGCTGTGGT GGGTTA	Tnfaip3	CAC TTTGTACCCCTGGT GAC	CCTACCC CCGT TCTGTAA
Tbx21	CAAAGTTCAACCAGCA CCAGAC	CCACGGTGAAGGACAG GAA	Traf2	TCTGTCCC AATGATGGAT GCA	GCAGGAATGGGCA AAGTCC
Tnf	CAAATGGCCCTCCCTCT CATCA	GCTACAGGCTTGTCACTC GAA	Vav1	CCTCTGCA GCCGATTCC TTAA	GTGGGTATGCACA GAGAAACA
Tnfrsfla	AGCTTGTGTCCCCAA GGAAA	CGGACAGTCACTCACCA AGTA	Zap70	GTGTCCTCCTGAGATGTA TGCA	GTTC CGCATACGT TGTCCA
Tnfrsflb	AGTGCATGAGGCTGA GCAA	GCC TTGCATAGCACATT TCCA	Zeb2	GGCAA GGCC TTC AAGTA CAA	TGCAGTTTGGG CAT TCGTAA

8. 2× KAPPA HiFi HotStart ReadyMix PCR kit (KAPA Biosystems, Wilmington, MA, USA) (similar High Fidelity PCR ready mixes can be substituted).
9. E-Gel® 2%EX Gel (ThermoFisher Scientific, Waltham, MA, USA).
10. DNA Clean & Concentrator™-5 kit (Zymo Research, Irvine, CA, USA).
11. MiSeq Sequencing System (Illumina, San Diego, CA, USA).

3 Methods

3.1 Sample Preparation

General measures are taken to avoid RNA/DNA contamination and degradation of samples. This includes using only sterile, RNAase- and DNAase-free tips, tubes, and plates, frequent changing of gloves, and keeping a clean work environment. Unless otherwise noted samples are kept at 4 °C in between reactions.

3.1.1 Single-Cell Sorting

1. Harvest organs of interest from a mouse and produce single-cell suspensions using a 70 µm cell strainer and the piston of a 3 mL sterile syringe. Harvest and wash in sterile PBS. Spin cells down (1200 rpm (350×g), 5 min, 20 °C) (*see Note 7*).
2. Wash samples once in FACS buffer and lyse red blood cells (RBC) by incubating for 5 min in 5 mL RBC lysis buffer, followed by another wash in FACS buffer (*see Note 8*).
3. Incubate cells in Fc-block for 15 min at room temperature. Meanwhile, distribute the cells into a conical bottom 96-well plate. Extra wells are included for compensation of each color and a blank, in addition to a negative control (*see Note 9*).
4. Wash once in FACS buffer and stain samples with PE-labeled tetramers for 1 h at room temperature in darkness. The compensations and blank samples are incubated in FACS buffer only.
5. Wash the cells three times. Samples are stained with the following mix of antibodies: FITC anti-mouse CD3ε, APC/Cy7 anti-mouse CD4, APC anti-mouse/human CD11b, APC anti-mouse/human CD45R/B220, APC anti-mouse CD49b, and APC anti-mouse CD8a. The compensations are stained with FITC anti-mouse CD3ε, APC/Cy7 anti-mouse CD4, APC anti-mouse/human, CD45R/B220, and PE anti-mouse CD4, respectively. The blank is resuspended in FACS buffer only. Incubate the cells for 30 min at 4 °C, in darkness.
6. Wash the cells twice in FACS buffer, and transfer each well to 500 µL FACS buffer in 5 mL polystyrene tubes.

- Sort single cells directly into a 96-well PCR plate containing 5 μL per well of RT Mix Solution 1 (*see* **Notes 10** and **11**). Cells are sorted by gating an APC-negative and FITC-, APC/Cy7-, and PE-positive population. Seal the plates immediately after sorting with adhesive plate seals, and briefly spin on a prechilled centrifuge. Plates can be frozen on dry ice and stored at -80°C .

3.1.2 Reverse Transcription and Pre-amplification

The following protocols are based on Fluidigm's user guide "Real-Time PCR Analysis". All the reactions are performed in a Thermal Cycler and the sample plate briefly spun before and after each reaction. Due to the small volumes and repetitive pipetting, only electronic pipettes are used.

- Thaw the sample plate on ice and denature the RNA by incubating for 90 s at 65°C , followed by immediate chilling on ice for 5 min.
- Add 1 μL of RT Mix Solution 2 to each well and do reverse transcription under the following conditions: 25°C , 5 min; 50°C , 30 min; 55°C , 25 min; 60°C , 5 min; 70°C , 10 min; final hold at 4°C .
- Pre-amplify the target cDNA by adding 9 μL STA Reaction Mix to each well and run the following reaction: 95°C , 10 min; 96°C , 5 s and 60°C , 4 min (20 cycles); final hold at 4°C (*see* **Note 12**).
- Remove unincorporated primers by adding 6 μL exonuclease to each well and vortex the plate for 20 s before spinning and placing it in the Thermal Cycler: 37°C , 30 min; 80°C , 15 min; final hold on 4°C .
- Dilute the final products by adding 54 μL $1\times$ DNA suspension buffer (fivefold dilution of the STA reaction), and store the STA pre-amplified sample plate at -20°C .

3.2 qPCR

All Fluidigm products must be warmed to room temperature before usage.

- Prime a 96.96 Dynamic Array IFC by injecting 150 μL control line fluid into each of the two accumulators and place the chip in an IFC Controller HX. Run the script "Prime (136 \times)" (*see* **Note 13**).
- Prepare a sample plate by distributing 4.4 μL freshly made Sample Pre-Mix Solution into each well of a 96-well plate. Using a multi-channel pipette transfer 3.6 μL from each well of the STA pre-amplified sample plate into the Sample Pre-Mix (*see* **Note 14**).
- Similarly, prepare an assay plate by distributing 9.5 μL freshly made Assay Mix Solution into each well of a 96-well plate.

Using a multi-channel pipette transfer 0.5 μL from each well of the qPCR Primer Plate into the Assay Mix Solutions.

4. Place adhesive seals and vortex each plate for 20 s and centrifuge for 30 s.
5. Carefully load 5 μL of each sample- and assay well into the respective inlets in the primed IFC. Place in the IFC Controller HX and run the script “Load Mix (136 \times)” (*see Note 15*).
6. Place the loaded IFC in the BioMark HD. Using the BioMark Data Collection software, select the protocol “GE 96 \times 96 Fast PCR + Melt v2.pcl” to run a thermal mix and qPCR under the following conditions: 70 $^{\circ}\text{C}$, 40 min; 60 $^{\circ}\text{C}$, 30 s; 95 $^{\circ}\text{C}$, 1 min; 96 $^{\circ}\text{C}$, 5 s and 60 $^{\circ}\text{C}$, 20 s (30 cycles); 60 $^{\circ}\text{C}$, 3 s; 60–95 $^{\circ}\text{C}$, 1 $^{\circ}\text{C}/3$ s (*see Note 16*).
7. Analyze the data using Real-Time PCR Analysis 4.0.1 software, and visualize results using SINGuLAR Analysis Toolset 3.0.

3.3 TCR Sequencing

3.3.1 Access Array

1. Prime a 48.48 Access Array IFC by injecting 300 μL control line fluid into each of the two accumulators, and add 500 μL 1 \times Access Array Harvest solution to wells H1–H3, and 500 μL 1 \times Access Array Hydration Reagent v2 to well H4. Place the chip in an IFC Controller AX (pre-PCR) and run the script “Prime (151 \times)” (*see Note 13*).
2. Make a sample plate in half of a 96-well plate by distributing 7.2 μL freshly made Sample Pre-Mix Solution per well. With a multi-channeled pipette, transfer 2 μL sample from each well of one half of the STA pre-amplified sample plate, and 0.8 μL from the Barcoded Internal TCR 3' Primer Plate (*see Notes 14 and 17*).
3. Place adhesive seals and vortex the sample plate and the Internal TCR 5' Primer Plate (20 \times) for 20 s and centrifuge for 30 s.
4. Carefully load 4 μL from each well of the sample plate and the 5' Primer Plate (20 \times) into the respective inlets of the primed IFC. Place the chip in the same IFC Controller AX as used for priming and run the script “Load Mix (151 \times)” (*see Note 15*).
5. Transfer the IFC to the BioMark HD, and using the BioMark Data Collection software, start the protocol “AA No I 48 \times 48 Standard v1” to run a thermal mix and PCR under the following conditions: 50 $^{\circ}\text{C}$, 2 min; 70 $^{\circ}\text{C}$, 20 min; 95 $^{\circ}\text{C}$, 10 min; 95 $^{\circ}\text{C}$, 15 s, 60 $^{\circ}\text{C}$, 30 s, and 72 $^{\circ}\text{C}$, 60 s (10 cycles); 95 $^{\circ}\text{C}$, 15 s, 80 $^{\circ}\text{C}$, 30 s, 60 $^{\circ}\text{C}$, 30 s, and 72 $^{\circ}\text{C}$, 60 s (2 cycles); 95 $^{\circ}\text{C}$, 15 s, 60 $^{\circ}\text{C}$, 30 s, and 72 $^{\circ}\text{C}$, 60 s (8 cycles); 95 $^{\circ}\text{C}$, 15 s, 80 $^{\circ}\text{C}$, 30 s, 60 $^{\circ}\text{C}$, 30 s, and 72 $^{\circ}\text{C}$, 60 s (2 cycles); 95 $^{\circ}\text{C}$, 15 s, 60 $^{\circ}\text{C}$, 30 s, and 72 $^{\circ}\text{C}$, 60 s (8 cycles); 95 $^{\circ}\text{C}$, 15 s, 80 $^{\circ}\text{C}$, 30 s, 60 $^{\circ}\text{C}$, 30 s, and 72 $^{\circ}\text{C}$, 60 s (5 cycles); 72 $^{\circ}\text{C}$, 3 min; final hold at 10 $^{\circ}\text{C}$.

6. After the PCR is finished, remove the remaining fluid from wells H1–H4 and replace with 600 μ L 1 \times Access Array Harvest Solution. Add 2 μ L 1 \times Access Array Harvest Solution into each of the sample inlets.
7. Place the IFC in a different IFC Controller AX (post-PCR) and harvest the amplification products back into the sample inlets by using the script “Harvest v5 (151 \times).”
8. Transfer the products from each sample inlet and combine into an Eppendorf tube and store at -20°C .

3.3.2 Library Construction and Illumina Sequencing

1. Clean up the pooled DNA amplicons by using AMPure XP magnetic bead isolation kit. Add 1.8 \times volume of the magnetic bead solution (based on volume of sample pool), allow nucleic acid amplicons to bind to the magnetic beads for 10 min, then place tube in a magnetic separator and wash the beads with 80% ethanol. Remove ethanol from the tube and repeat wash step, for a total of two washes. Allow magnetic beads to dry with the lid of the tube open at room temperature for 10 min. Remove the tube from the magnetic separator and add 30 μ L of nuclease-free water, wait 2 min for the amplicon products to elute off the beads, and then place the tube in the magnetic separator and remove the 30 μ L nuclease-free water containing the DNA libraries and transfer it to a fresh microfuge tube.
2. Analyze cleaned up DNA amplicon products for length distribution using an Agilent 2100 Bioanalyzer instrument and High Sensitivity dsDNA chip.
3. Quantitate cleaned up DNA amplicon products with a Qubit fluorometer and dsDNA high sensitivity assay.
4. Using quantitation values from the Qubit, take 10 ng of the DNA amplicons (*from step 1*), into the NEBNext Ultra DNA Library Prep Kit for Illumina; follow the manufacturer’s guidelines for library preparation (during the protocol, no magnetic bead size-selection is performed and PCR amplification is done with 15 cycles).
5. After PCR amplification, the DNA libraries are cleaned up with AmpureXP magnetic beads (*see step 1* above) and the cleaned up products are purified using a 2% E-Gel EX agarose gel. DNA products in the range of 270–490 bp (*see Note 18*) are excised and isolated from the gel using Zymo Agarose Dissolving Buffer (ADB) and DNA Clean & ConcentratorTM-5 kit (*see Note 19*). Elute the DNA library products in 20 μ L nuclease-free water.
6. The DNA libraries are diluted to the appropriate concentrations, and loaded onto the MiSeq system following the manufacturer’s guidelines (*see Note 20*).

Table 2
External primers for pre-amplification (TCR sequencing)

Target	Forward (TRAV)	Target	Forward (TRBV)
TRAV1	GGTTATCCTGGTACCAGCA	TRBV1	TACCACGTGGTCAAGCTG
TRAV2	CATCTACTGGTACCGACAGG	TRBV2	CAGTATCTAGGCCACAATGC
TRAV3	GGCGAGCAGGTGGAG	TRBV3	CCCAAAGTCTTACAGATCCC
TRAV4	TCTGSTCTGAGATGCAAITTT	TRBV4	GACGGCTGTTTTCCAGAC
TRAV5-1/5-4(D)	GGCTACTTCCCTTGGTATAAGCAAGA	TRBV5	GGTATAAACAGAGCGGCTGAG
TRAV6-1/6-2	CAGATGCAAGGTCAAGTGAC	TRBV12	GGGGTTGTCCAGTCTCC
TRAV6-3/6-4(D)	AAGGTCCACAGCTCCTTC	TRBV13	GCTGCAGTCACCCAAAAG
TRAV6-5/6-7(D)	GTTCTGGTATGTGCAGTATCC	TRBV14	GCAGTCCCTACAGGAAGGG
TRAV6-6	AGATTCCGTGACTCAAACAG	TRBV15	GAGTTAACCAGACACCCAG
TRAV7	AGAAGGTRCAGCAGAGCCCCAGAATC	TRBV16	CCTAGGCACAAGGTGACAG
TRAV8	GAGCRICCSAGGGGTG	TRBV17	GAAGCCAAAACCAAGCAC
TRAV9	CCAGTGGTTCAAGGAGTG	TRBV19	GATTGGTCAGGAAGGGC
TRAV10/10a(D)	AGAGAAAGGTCGAGCAACAC	TRBV20	GGATGGAGTGTCAAGCTG
TRAV11	AAGACCCAAGTGGAGCAG	TRBV23	CTGCAGTTACACAGAAGCC
TRAV12	TGACCCAGACAGAAGGC	TRBV24	CAGACTCCACGATACCTGG
TRAV13	TCCTTGGTTCTGCAGG	TRBV26	GGTGAAAGGGCAAGGAC

TRAV14	GCAGCAGGTGAGACAAAG	TRBV29	GCTGGAATGTGGACAGG
TRAV15	CASCTTYTTAGTGGAGAGATGG	TRBV30	CCTCCTCTACCACAAAAGCC
TRAV16	GTACAAAGCAAACAGCAAAGTG	TRBV31	CTAACCTCTACTGGTACTGGCAG
TRAV17	CAGTCCGTGGACCAGC		
TRAV18	AACGGCTGGAGCAGAG		
TRAV19	GCAAGTTAAACAAAAGCTCTCC		
TRAV21	GTGCACCTTGCCCTGTAGC		
	Reverse (TRAC)		Reverse (TRBC)
TRAC-rev.	GGCATCACAGGGAAACG	TRBC-rev.	CCAGAAGGTAGCAGAGACCC

Table 3
Internal 5' primers for access array (TCR sequencing)

Target	Forward (TRAV)	Target	Forward (TRBV)
TRAV1	CTCCACATTCCTGAGCC	TRBV1	GTATCCCTGGATGAGCTG
TRAV2	ACTCTGAGCCTGCCCT	TRBV2	GGACAATCAGACTGCCTC
TRAV3	GCCCTCCTCACCTGAG	TRBV3	GATATGGGGCAGATGGTG
TRAV4	GGITMAGGAACAAGGAGAAT	TRBV4	CAGGTGGAAATGAAAGTG
TRAV5-1/5-4(D)	ATYCGTTCAAATATGGAAAGAAA	TRBV5	GCCAGAGCTCATGTTTCTC
TRAV6-1/6-2	GGAGAAGGTCCACAGCTC	TRBV12	CCAGCAGATTCTCAGTCC
TRAV6-3/6-4(D)	CAACTGCCAACAAACAAGG	TRBV13	GTACTGGTATCGGCAGGAC
TRAV6-5/6-7(D)	TCCTTCCACTTGCAGAAAAG	TRBV14	GGTATCAGCAGCCCCAGAG
TRAV6-6	ACGGCTGGCCAGAAG	TRBV15	GTGTGAGCCAGTTTCAGG
TRAV7	CAKGRCYTCYYTCAACTGCAC	TRBV16	GAAGCAACTCTGTGGTGTG
TRAV8	AGAGCCACCCTTGACAC	TRBV17	GAACAGGGAAGCTGACAC
TRAV9	GCTTYGAGGCTGAGTTCAG	TRBV19	GGTACCGACAGGATTCAG
TRAV10/10a(D)	CTACACTGAGTGTTCGAGAGG	TRBV20	GCTTGGTATCGTCAATCG
TRAV11	AACAGGACACAGGCAAAAG	TRBV23	GCCAGGAAGCAGAGATG
TRAV12	GGTTCCACGCCACTC	TRBV24	GCACACTGCCITTTTACTGG
TRAV13	TGCAGGAGGGGAGA	TRBV26	GAGGTGTATCCCTGAAAAGG

TRAV14	CTCTGACAGTCTGGGAAGG	TRBV29	GTACTGGTATCGACAAGACCC
TRAV15	AYTCTGTAGTCTTCCAGAAATCAC	TRBV30	GGACATCTGTCAAAGTGGC
TRAV16	ATTATTCTCTGAACITTCAGAAGC	TRBV31	CTGTTGGCCAGGTAGAGTC
TRAV17	TATGAAAGGAGCCTCCCTG		
TRAV18	CAAGATTTCACCGCACG		
TRAV19	GCTGACTGTTCAAGAGGGA		
TRAV21	AATAGTATGGCTTTCCTGGC		

Table 4
Internal barcoded 3' primers for access array (TCR sequencing)

Name	Barcoded Reverse (TRAC)	Name	Barcoded Reverse (TRBC)
BC1TRAC	ATCACGGCACATTGATTTGGGAGTC	BC1TRBC	ATCACGGGGTAGCCTTTTTGTTTG
BC2TRAC	CGATGTGCACATTGATTTGGGAGTC	BC2TRBC	CGATGTGGGTAGCCTTTTTGTTTG
BC3TRAC	TTAGGCGCACATTGATTTGGGAGTC	BC3TRBC	TTAGGCGGGTAGCCTTTTTGTTTG
BC4TRAC	TGACCAGCACATTGATTTGGGAGTC	BC4TRBC	TGACCAGGGTAGCCTTTTTGTTTG
BC5TRAC	ACAGTGGCACATTGATTTGGGAGTC	BC5TRBC	ACAGTGGGGTAGCCTTTTTGTTTG
BC6TRAC	GCCAAATGCACATTGATTTGGGAGTC	BC6TRBC	GCCAAATGGGTAGCCTTTTTGTTTG
BC7TRAC	CAGATCGCACATTGATTTGGGAGTC	BC7TRBC	CAGATCGGGTAGCCTTTTTGTTTG
BC8TRAC	ACTTGAGCACATTGATTTGGGAGTC	BC8TRBC	ACTTGAGGGTAGCCTTTTTGTTTG
BC9TRAC	GATCAGGCACATTGATTTGGGAGTC	BC9TRBC	GATCAGGGGTAGCCTTTTTGTTTG
BC10TRAC	TAGCTTGACACATTGATTTGGGAGTC	BC10TRBC	TAGCTTGGGTAGCCTTTTTGTTTG
BC11TRAC	GGCTACGCACATTGATTTGGGAGTC	BC11TRBC	GGCTACGGGTAGCCTTTTTGTTTG
BC12TRAC	CTTGTAGCACATTGATTTGGGAGTC	BC12TRBC	CTTGTAGGGTAGCCTTTTTGTTTG
BC13TRAC	AGTCAAGCACATTGATTTGGGAGTC	BC13TRBC	AGTCAAGGGTAGCCTTTTTGTTTG
BC14TRAC	AGTCCGCACATTGATTTGGGAGTC	BC14TRBC	AGTCCGGGTAGCCTTTTTGTTTG
BC15TRAC	ATGTCAGCACATTGATTTGGGAGTC	BC15TRBC	ATGTCAGGGTAGCCTTTTTGTTTG
BC16TRAC	CCGTCCGCACATTGATTTGGGAGTC	BC16TRBC	CCGTCCGGGTAGCCTTTTTGTTTG
BC17TRAC	GTAGAGGCACATTGATTTGGGAGTC	BC17TRBC	GTAGAGGGGTAGCCTTTTTGTTTG
BC18TRAC	GTCCGGCACATTGATTTGGGAGTC	BC18TRBC	GTCCGGGGTAGCCTTTTTGTTTG
BC19TRAC	GTGAAAGCACATTGATTTGGGAGTC	BC19TRBC	GTGAAAGGGTAGCCTTTTTGTTTG
BC20TRAC	GTGGCCGCACATTGATTTGGGAGTC	BC20TRBC	GTGGCCGGGTAGCCTTTTTGTTTG

BC21TRAC	GTTTCGGCACATTGATTTGGGAGTC	BC21TRBC	GTTTCGGGGTAGCCTTTTGTTTGTTTG
BC22TRAC	CGTACGGCACATTGATTTGGGAGTC	BC22TRBC	CGTACGGGGTAGCCTTTTGTTTGTTTG
BC23TRAC	GAGTGGGCACATTGATTTGGGAGTC	BC23TRBC	GAGTGGGGGTAGCCTTTTGTTTGTTTG
BC24TRAC	GGTAGCGCACATTGATTTGGGAGTC	BC24TRBC	GGTAGCGGGTAGCCTTTTGTTTGTTTG
BC25TRAC	ACTGATGCACATTGATTTGGGAGTC	BC25TRBC	ACTGATGGGTAGCCTTTTGTTTGTTTG
BC26TRAC	ATGAGCGCACATTGATTTGGGAGTC	BC26TRBC	ATGAGCGGGTAGCCTTTTGTTTGTTTG
BC27TRAC	ATTCCCTGCACATTGATTTGGGAGTC	BC27TRBC	ATTCCCTGGGTAGCCTTTTGTTTGTTTG
BC28TRAC	CAAAAGGCACATTGATTTGGGAGTC	BC28TRBC	CAAAAGGGGTAGCCTTTTGTTTGTTTG
BC29TRAC	CAACTAGCACATTGATTTGGGAGTC	BC29TRBC	CAACTAGGGTAGCCTTTTGTTTGTTTG
BC30TRAC	CACCGGGCACATTGATTTGGGAGTC	BC30TRBC	CACCGGGGTAGCCTTTTGTTTGTTTG
BC31TRAC	CACGATGCACATTGATTTGGGAGTC	BC31TRBC	CACGATGGGTAGCCTTTTGTTTGTTTG
BC32TRAC	CACTCAGCACATTGATTTGGGAGTC	BC32TRBC	CACTCAGGGTAGCCTTTTGTTTGTTTG
BC33TRAC	CAGGCGGCACATTGATTTGGGAGTC	BC33TRBC	CAGGCGGGGTAGCCTTTTGTTTGTTTG
BC34TRAC	CATGGCGCACATTGATTTGGGAGTC	BC34TRBC	CATGGCGGGTAGCCTTTTGTTTGTTTG
BC35TRAC	CATTTTGCACATTGATTTGGGAGTC	BC35TRBC	CATTTTGGGTAGCCTTTTGTTTGTTTG
BC36TRAC	CCAACAGCACATTGATTTGGGAGTC	BC36TRBC	CCAACAGGGTAGCCTTTTGTTTGTTTG
BC37TRAC	CGGAATGCACATTGATTTGGGAGTC	BC37TRBC	CGGAATGGGTAGCCTTTTGTTTGTTTG
BC38TRAC	CTAGCTGCACATTGATTTGGGAGTC	BC38TRBC	CTAGCTGGGTAGCCTTTTGTTTGTTTG
BC39TRAC	CTATACGCACATTGATTTGGGAGTC	BC39TRBC	CTATACGGGTAGCCTTTTGTTTGTTTG
BC40TRAC	CTCAGAGCACATTGATTTGGGAGTC	BC40TRBC	CTCAGAGGGTAGCCTTTTGTTTGTTTG

(continued)

Table 4
(continued)

Name	Barcoded Reverse (TRAC)	Name	Barcoded Reverse (TRBC)
BC41TRAC	GACGACGCACATTGATTTGGGAGTC	BC41TRBC	GACGACGGGTAGCCTTTTGTTTGTTTG
BC42TRAC	TAATCGGCACATTGATTTGGGAGTC	BC42TRBC	TAATCGGGGTAGCCTTTTGTTTGTTTG
BC43TRAC	TACAGCGCACATTGATTTGGGAGTC	BC43TRBC	TACAGCGGTAGCCTTTTGTTTGTTTG
BC44TRAC	TATAATGCACATTGATTTGGGAGTC	BC44TRBC	TATAATGGGTAGCCTTTTGTTTGTTTG
BC45TRAC	TCATTCGCACATTGATTTGGGAGTC	BC45TRBC	TCATTCGGGTAGCCTTTTGTTTGTTTG
BC46TRAC	TCCCGAGCACATTGATTTGGGAGTC	BC46TRBC	TCCCGAGGTAGCCTTTTGTTTGTTTG
BC47TRAC	TCGAAGGCACATTGATTTGGGAGTC	BC47TRBC	TCGAAGGGGTAGCCTTTTGTTTGTTTG
BC48TRAC	TCGGCAGCACATTGATTTGGGAGTC	BC48TRBC	TCGGCAGGTAGCCTTTTGTTTGTTTG

3.4 Data Analysis

1. TCR sequencing analysis is always started by the identification of bar codes.
2. TCR chain identity is determined using the IMGT database (<http://www.imgt.org>).

4 Notes

1. Biotinylated MHC molecules can be produced in house (our case) or obtained from the NIH tetramer core facility.
2. Include primer pairs for all targets of both qPCR and TCR α and β sequencing. We combine the same 96 target-specific primer pairs as used during the qPCR (*see* Table 1), with the 42 external primer pairs for TCR sequencing (*see* Table 2), the latter being part of a nested PCR.
3. The sequences of primers for TCR sequencing were published by Thomas et al. [8].
4. The gene expression profiling can be carried out for any gene of interest. Our particular panel includes seven categories: cell surface receptors, chemokine receptors, cytokines, transcription factors, interferon response, metabolism, and signaling molecules. Primers pairs were selected from the Fluidigm catalog of validated reagents (DELTAgene Assays) (*see* Table 1).
5. Two separate IFCs are needed, one for pre- and one for post-PCR.
6. If less than 48 5' primers: use 1 \times DNA Suspension Buffer (TEKnova) in the remaining wells. If more than 48 5' primers: combine multiple primers per well.
7. The timing of tissue harvesting \rightarrow cell staining \rightarrow cell sorting \rightarrow reverse transcription should be as tight as possible to avoid cell death and poor quality cDNA synthesis.
8. Lysis of RBCs is done for splenocytes only.
9. For compensations and the negative control we use splenocytes, as these are the most abundant of our samples. Before distribution, 10 μ L is removed from each sample to calculate total cell numbers for records.
10. qPCR experiments should include a minimum of four blank wells and, if possible, wells with 5, 10, and 50 cells.
11. We sort into a hard-shelled, deep welled PCR plate to avoid cross contamination in the downstream processing.
12. If the sample plate is to be used only for access array and not qPCR, it can be an advantage to increase to 25 cycles.
13. The primed chip must be loaded within an hour of finishing priming.

14. A minimum overage of 3 μL /well is helpful to avoid loading air bubbles.
15. The loaded chip must immediately be transferred to the Biomark upon finishing.
16. Turn on the Biomark at least 20 min prior to use to allow the camera to cool to 4 $^{\circ}\text{C}$. Use scotch tape to gently remove potential dust from the chip surface.
17. Pairing of TCR α and β chains is made possible by using internal barcoded 3' primers binding TRAC and TRBC (*see* Table 4). As access array allows for up to 48 samples, we designed 48 6 bp barcodes that were added to the TRAC and TRBC 3' primers. By loading the barcode-based 3' primer pairs together with the samples (instead of with the 5' primers as suggested in Fluidigm's protocol for Access Array), the TCR α and β sequences from the same cell will share the same barcode.
18. Sizes based on primers from Tables 3 and 4.
19. 4 \times volume of ADB should be used to dissolve the gel and the mixture should be incubated at 37 $^{\circ}\text{C}$ for 5–10 min, until the gel is dissolved.
20. A 2 \times 300 sequencing run is performed based on estimated 5 M reads per sample.

References

1. Flatz L, Roychoudhuri R, Honda M, Filali-Mouhim A, Goulet JP, Kettaf N, Lin M, Roederer M, Haddad EK, Sekaly RP, Nabel GJ (2011) Single-cell gene-expression profiling reveals qualitatively distinct CD8 T cells elicited by different gene-based vaccines. *Proc Natl Acad Sci U S A* 108:5724–5729. PMID: 3078363
2. Narsinh KH, Sun N, Sanchez-Freire V, Lee AS, Almeida P, Hu S, Jan T, Wilson KD, Leong D, Rosenberg J, Yao M, Robbins RC, Wu JC (2011) Single cell transcriptional profiling reveals heterogeneity of human induced pluripotent stem cells. *J Clin Invest* 121:1217–1221. PMID: 3049389
3. Yoshida K, Corper AL, Herro R, Jabri B, Wilson IA, Teyton L (2010) The diabetogenic mouse MHC class II molecule I-Ag7 is endowed with a switch that modulates TCR affinity. *J Clin Invest* 120:1578–1590. PMID: 2860908
4. Foldy C, Darmanis S, Aoto J, Malenka RC, Quake SR, Sudhof TC (2016) Single-cell RNAseq reveals cell adhesion molecule profiles in electrophysiologically defined neurons. *Proc Natl Acad Sci U S A* 113:E5222–E5231. PMID: PMC5024636
5. Brouzes E (2012) Droplet microfluidics for single-cell analysis. *Methods Mol Biol* 853:105–139
6. Yin H, Marshall D (2012) Microfluidics for single cell analysis. *Curr Opin Biotechnol* 23:110–119
7. Newell EW, Davis MM (2014) Beyond model antigens: high-dimensional methods for the analysis of antigen-specific T cells. *Nat Biotechnol* 32:149–157. PMID: PMC4001742
8. Dash P, McClaren JL, Oguin TH 3rd, Rothwell W, Todd B, Morris MY, Becksfort J, Reynolds C, Brown SA, Doherty PC, Thomas PG (2011) Paired analysis of TCR α and TCR β chains at the single-cell level in mice. *J Clin Invest* 121:288–295. PMID: PMC3007160

Investigate Global Chromosomal Interaction by Hi-C in Human Naive CD4 T Cells

Xiangzhi Meng, Nicole Riley, Ryan Thompson, and Siddhartha Sharma

Abstract

Hi-C is a methodology developed to reveal chromosomal interactions from a genome-wide perspective. Here, we described a protocol for generating Hi-C sequencing libraries in resting and activated human naive CD4 T cells to investigate activation-induced chromatin structure re-arrangement in T cell activation followed by a section reviewing the general concepts of Hi-C data analysis.

Key words Chromosomal interaction, Cell fixation, Endo-restricted enzyme digestion, De novo ligation, Biotin-streptavidin, High-throughput sequencing

1 Introduction

In eukaryotic cells, the basic unit of chromatin is the 147 bp of genomic DNA wrapping around an octamer of proteins called histones. This basic unit of DNA is called a nucleosome [1]. Nucleosomes are then packaged into higher levels of compaction. The dynamics of chromatin conformation has been shown in multiple cell types and this highly ordered structure is demonstrated to play important roles in transcription in multiple biological processes [2–4]. Chromatin interactions are mediated by mediator, cohesin, as well as by other transcription factors and co-factors. Chromosome Conformation Capture (3C) [5] and three-dimensional DNA fluorescence in situ hybridization (3D-FISH) [6] had been developed to examine chromatin interactions at certain loci, but an approach to examining global, genome-wide interactions was unavailable until 2009 when Hi-C was developed. Hi-C is a methodology aimed to comprehensively investigate chromatin interactions in nuclei, and was first described by Lieberman-Aiden et al. [7]. Using similar approaches as 3C, Hi-C also starts with crosslinking cells with formaldehyde to preserve the chromatin structure. And in the following steps crosslinked chromatin is digested with endo-restricted enzymes, and then the digestion

product is re-ligated with biotin-labeled nucleotides. Thus, the de novo ligation sites, where the information of DNA fragments that are physically linked together is restored, can then be purified by streptavidin-biotin interactions. By deep sequencing these ligation products and subsequent computational analysis, the genome-wide physical interactions between chromatin sites can be resolved and the 3D organization of the genome could be simulated [8, 9]. The high resolution of deep sequencing enables investigators to survey not only the spatial organization of chromatin and its biophysical properties but also the biological functions of specific chromatin loci [10–12]. This technique provides unprecedented insight in the regulation of the genome in normal conditions and in disease states.

The protocol described here is derived from a *in situ* Hi-C method demonstrated by Rao et al. [10], and optimized for human primary naive CD4 T cells.

2 Materials

2.1 Crosslinking

1. Culture media: RPMI 1640 supplemented with 100 U/mL Penicillin, 100 µg/mL Streptomycin, and 10% Fetal bovine serum (FBS).
2. 16% formaldehyde (methanol free).
3. Phosphate-buffered saline (PBS): 137 mM NaCl, 2.7 mM KCl, 10 mM Na₂HPO₄, 1.8 mM KH₂PO₄.
4. 3 M Tris-HCl (pH 8.0).

2.2 Restriction Digestion

1. Digestion buffer (10×): 500 mM Potassium Acetate, 200 mM Tris-Acetate, 100 mM Magnesium Acetate, 1 mg/mL Bovine serum albumin (BSA), pH 7.9 at 25 °C.
2. 10% Sodium dodecyl sulfate (SDS) solution.
3. 20% Triton X-100 solution.
4. Hind III endo-restricted enzyme (20 units/µL).
5. Buffer 1: 11 mM HEPES, 11 mM KCl, 1.65 mM MgCl₂, 11% Glycerol, 0.55% NP-40, 1.1% Triton X-100.

2.3 Re-ligation

1. 0.4 mM biotin-14-dCTP.
2. 10 mM dATP.
3. 10 mM dGTP.
4. 10 mM dTTP.
5. DNA Polymerase I, Large (Klenow) Fragment (5 units/µL).
6. T4 ligase buffer (10×): 500 mM Tris-HCl, 100 mM MgCl₂, 10 mM ATP, 10 mM DTT, pH 7.5 at 25 °C.

7. T4 DNA Ligase (2000 units/ μ L).
8. Proteinase K (20 mg/mL).

2.4 Digestion and Ligation Efficiency Assessment

1. Phenol-Chloroform-Isoamyl Alcohol (25:24:1, pH 8.3).
2. 1% agarose gel.

2.5 DNA Extraction, DNA Shearing, and Size Selection

1. 3 M sodium acetate (pH 5.0).
2. 70% ethanol (w/w).
3. Tris buffer (1 \times): 10 mM Tris-HCl (pH 8.0).
4. Covaris instrument and Covaris microTUBE (with AFA fiber).
5. AMPure XP beads.
6. 80% ethanol.

2.6 Biotin Pull-Down and Preparation of Illumina Sequencing

1. 10 mg/mL streptavidin beads.
2. Wash buffer: 5 mM Tris-HCl (pH 7.5), 0.5 mM EDTA, 1 M NaCl, 0.50% Tween-20.
3. Binding buffer (2 \times): 10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 2 M NaCl.
4. 25 mM dNTP mix.
5. T4 PNK (10 Units/ μ L).
6. T4 DNA Polymerase I (3 Units/ μ L).
7. DNA polymerase I, Large (Klenow) Fragment (5 Units/ μ L).
8. dATP attachment buffer: 50 mM NaCl, 10 mM Tris-HCl, 10 mM MgCl₂, 1 mM DTT, pH 7.9 at 25 °C.
9. Klenow exo minus (5 units/ μ L).
10. Quick ligation buffer: 66 mM Tris-HCl, 10 mM MgCl₂, 1 mM DTT, 1 mM ATP, 6% Polyethylene glycol (PEG 6000), pH 7.6 at 25 °C.
11. DNA quick ligase.
12. 25 μ M Illumina indexed adapter.
13. Illumina Truseq primers 1 and 2.
14. KAPA HiFi PCR master mix.
15. Qubit dsDNA High Sensitivity Assay Kit.
16. TapeStation and D1000 DNA tape kit.

3 Methods

3.1 Crosslinking

Isolate peripheral blood mononuclear cells (PBMC) from healthy donor's blood by centrifugation through a histopaque (Sigma) gradient with deceleration break-off. Then purify naive CD4 T cells via the negative selection naive CD4 T cell isolation kit.

1. Resuspend 10 million naive CD4 T cells in 5 mL warm culture medium (*see Note 1*).
2. Add 312 μL of 16% formaldehyde (methanol free) (final concentration is 1%) (*see Note 2*) and rotate at room temperature (room temperature) for 5 min.
3. Add 1732 μL of 3 M Tris-HCl (pH 8.0) (final concentration is 750 mM) to quench formaldehyde, rotate at room temperature for 5 min.
4. Spin down the tube at 4 °C for 10 min at 1800 rpm ($300\times g$).
5. Remove the supernatant, wash the pellet with 10 mL cold PBS, spin down at 4 °C at 1800 rpm ($300\times g$) for 5 min.
6. Remove the supernatant and snap freeze the cell pellet in liquid nitrogen for 5 min.
7. Store the pellet in -80 °C for future use.

3.2 Restriction Digestion

1. Thaw the pellet on ice, meanwhile prepare the cell lysis buffer by combining 450 μL of ice-cold Buffer 1 with 50 μL of 10 \times proteinase inhibitor solution. Resuspend the pellet in the lysis buffer and rotate for 30 min in a cold room.
2. Centrifuge the tube at 5000 rpm ($2400\times g$) for 10 min at 4 °C, remove the supernatant, and resuspend the pellet in 500 μL 1.2 \times digestion buffer (combine 60 μL of 10 \times digestion buffer with 440 μL of H_2O).
3. Add 15 μL of 10% SDS (final concentration is 0.3%), mix up, and incubate at 62 °C for 10 min, followed by 37 °C for 50 min, both while shaking at 900 rpm.
4. To quench SDS, add 50 μL of 20% Triton X-100 (final concentration is 2%) to the reaction and incubate at 37 °C for 60 min, shaking at 900 rpm.
5. Bring the tube to room temperature and take a 50 μL aliquot for the undigested control (store in -20 °C).
6. To the rest, add 200 U of HindIII restriction enzyme (10 μL of 20 Units/ μL) and incubate overnight at 37 °C with shaking at 900 rpm.

3.3 Re-ligation and Reverse Crosslinking

1. The next morning, inactivate the restriction enzyme, incubate the tube at 62 °C for 20 min while shaking at 900 rpm
2. Cool down the tube to room temperature and add 50 μL of the following fill-in master mix (*see Note 3*):
 - (a) 37 μL of 0.4 mM biotin-14-dCTP.
 - (b) 1.5 μL of 10 mM dATP.
 - (c) 1.5 μL of 10 mM dGTP.
 - (d) 1.5 μL of 10 mM dTTP.
 - (e) 8 μL of 5 U/ μL DNA Polymerase I, Large (Klenow) Fragment.

3. Mix by pipetting and incubate at 37 °C for 1.5 h with shaking at 300 rpm.
4. Take a 50 µL aliquot for digested-unligated control and store it at -20 °C.
5. Transfer the rest of sample to a 15 mL tube and add 1672 µL of the following ligation master mix:
 - (a) 1326 µL of H₂O.
 - (b) 220 µL of 10× T4 DNA ligase buffer.
 - (c) 100 µL of 20% Triton X-100.
 - (d) 24 µL of 10 mg/mL Bovine Serum Albumin.
 - (e) 2 µL of 2000 Units/µL T4 DNA Ligase.
6. Mix by inverting the tube a couple of times and incubate the tube in a 16 °C water bath for 4 h.
7. Take a 50 µL aliquot for ligated control and store it at -20 °C.
8. To reverse the protein-DNA crosslinking, add 100 µL of 20 mg/mL proteinase K and 240 µL of 10% SDS, then incubate at the tube 55 °C for 30 min to degrade proteins. After that, add 260 µL of 5 M NaCl into the tube, mix up and split one tube into four 1.5 mL tubes (approximately 650 µL per 1.5 mL tube) and incubate these four tubes at 65 °C overnight with shaking at 900 rpm.
9. Meanwhile, thaw the undigested, digested-unligated and ligated controls out and bring the volume up to 100 µL with H₂O. Add 1 µL of RNase A into each tube and incubate at 37 °C for 30 min. Then add 2 µL of Proteinase K and 10 µL of 10% SDS to each control and incubate at 65 °C overnight with the Hi-C samples.

3.4 Digestion and Ligation Efficiency Assessment

1. To extract DNA from the controls, add 100 µL of phenol-chloroform-Isoamyl Alcohol (25:24:1) to each tube and vortex to completely mix.
2. Centrifuge the tubes at 10,000 rpm (9600×g) for 5 min.
3. Transfer the upper aqueous phase to a new tube and run the controls on a 1% agarose gel at 100 V for 20–30 min (*see Note 4*).

3.5 DNA Extraction, DNA Shearing, and Size Selection

1. To extract DNA from the Hi-C samples, add 650 µL of Phenol-Chloroform-Isoamyl Alcohol into each sample tube (1× volume), vortex to completely mix, and spin down at 10,000 rpm (9600×g) at room temperature for 5 min.
2. Transfer the upper phase (about 650 µL) to a new 2 mL tube and add 1.3 mL (2× volume) of pure ethanol and 65 µL (0.1× volume) of 3 M sodium acetate, then mix by inverting and incubate at -80 °C for 2 h.

3. Take out the samples from $-80\text{ }^{\circ}\text{C}$ freezer and centrifuge at max speed at $4\text{ }^{\circ}\text{C}$ for 15 min. Keep the tubes on ice after centrifugation and carefully remove the supernatant without disturbing the pellet.
4. Add $250\text{ }\mu\text{L}$ of 70% ice-cold ethanol into each tube and combine all four tubes of the same sample,
5. Centrifuge at max speed for 5 min at $4\text{ }^{\circ}\text{C}$.
6. Remove the entire supernatant and dry the pellet at room temperature for 10 min until the pellet becomes semi-transparent.
7. Dissolve the pellet in $145\text{ }\mu\text{L}$ of $1\times$ Tris buffer (10 mM Tris-HCl, pH 8.0).
8. To make the biotinylated DNA suitable for high-throughput sequencing using Illumina sequencers, shear the DNA to a size of 200 bp using the following parameters on Covaris S2:
 - (a) Vial: Covaris microTUBE (with AFA fiber).
 - (b) Volume: $130\text{ }\mu\text{L}$.
 - (c) Fill level: 12.
 - (d) Duty Cycle: 10%.
 - (e) Intensity: 5.
 - (f) Cycles/Burst: 200.
 - (g) Time: 180 s.
9. Transfer the sheared DNA to a new 1.5 mL tube. Bring the volume to $203\text{ }\mu\text{L}$ with H_2O . Take a $3\text{ }\mu\text{L}$ aliquot for sheared-control and examine the DNA fragment size on the D1000 tape (Tapestation) (*see Note 5*).
10. Warm a bottle of AMPure XP beads to room temperature.
11. Before using, vortex the beads to mix completely.
12. Add exactly $180\text{ }\mu\text{L}$ ($0.9\times$ volumes) of beads to each sample, mix well by pipetting at least ten times, and incubate at room temperature for 5 min.
13. Put the tube on a magnet and wait until the beads are completely attached to the magnetic side, then transfer the clear solution to a fresh tube, avoiding any beads (*see Note 6*).
14. Add exactly $50\text{ }\mu\text{L}$ ($0.25\times$ volumes) of fresh AMPure P beads to the transferred solution. Mix by pipetting at least ten times and incubate at room temperature for 5 min.
15. Separate the beads on a magnet and discard the clear solution (*see Note 6*).
16. While keeping the beads on the magnet, add $400\text{ }\mu\text{L}$ of 80% ethanol to the beads and discard the solution after 30 s. Wash with 80% ethanol for a second time.

17. Leave the beads on the magnet for 5 min to allow remaining ethanol to evaporate (do not over dry the beads, this could result in a reduced yield).
18. To elute the DNA, remove the tube from the magnet and add 301 μL of 1 \times Tris buffer and gently mix by pipetting. After incubating at room temperature for 5 min, separate the beads on a magnet and transfer the solution to a new 1.5 mL tube.
19. Take 1 μL of size selected DNA and verify the fragment size on the Tapestation (D1000 tape) (*see Note 6*).

3.6 Biotin Pull-Down and Preparation of Illumina Sequencing

Perform all the following steps in DNA low-binding tubes.

1. To prepare streptavidin beads for biotin pull-down, add 1 mL wash buffer to 150 μL of 10 mg/mL streptavidin beads, separate the beads on a magnet, and discard the solution. Repeat washing once more (total of 2 washes).
2. Remove the solution after the second wash and resuspend the beads in 300 μL of binding buffer. Add 300 μL of size-selected DNA to the beads and incubate at room temperature for 15 min with rotation.
3. Separate the beads on a magnet and discard the solution. Resuspend the beads in 600 μL of wash buffer and incubate at 55 $^{\circ}\text{C}$ for 2 min with shaking at 900 rpm. Retrieve the beads using a magnet and discard the supernatant.
4. Repeat the wash.
5. Resuspend the beads in 100 μL of 1 \times T4 DNA ligase buffer and retrieve the beads on a magnet. Discard the buffer.
6. Resuspend the beads in 100 μL of end-repair and biotin-remove master mix.
 - (a) 88 μL of 1 \times T4 DNA ligase buffer with 10 mM ATP.
 - (b) 2 μL of 25 mM dNTP mix.
 - (c) 5 μL of 10 U/ μL T4 PNK.
 - (d) 4 μL of 3 U/ μL T4 DNA Polymerase I.
 - (e) 1 μL of 5 U/ μL DNA polymerase I, Large (Klenow) Fragment.
7. Incubate at room temperature for 30 min, separate the beads on a magnet, and discard the supernatant.
8. Wash the beads by adding 600 μL of wash buffer and incubate at 55 $^{\circ}\text{C}$ for 2 min with mixing. Reclaim the beads and discard the supernatant.
9. Resuspend the beads in 100 μL dATP attachment buffer. Reclaim the beads and discard the supernatant.

10. Resuspend the beads in 100 μL of dATP attachment master mix:
 - (a) 90 μL of dATP attachment buffer.
 - (b) 5 μL of 10 mM dATP.
 - (c) 5 μL of 5 U/ μL Klenow exo minus.
11. Incubate the sample at 37 °C for 30 min. Separate the beads on a magnet and discard the solution.
12. Resuspend the beads with 600 μL of wash buffer and incubate at 55 °C for 2 min with mixing. Reclaim the beads with the magnet and discard the solution.
13. Wash the beads by resuspending the beads in 100 μL of quick ligation reaction buffer and resuspend the beads in 50 μL of quick ligation reaction buffer after wash.
14. Add 2 μL of DNA quick ligase plus 3 μL of a 25 μM Illumina indexed adapter, mix thoroughly.
15. Incubate at room temperature for 15 min, then separate the beads on a magnet and discard the solution.
16. Wash the beads by resuspending the beads in 600 μL of wash buffer.
17. Incubate the beads at 55 °C for 2 min with mixing. Reclaim the beads and remove the solution.
18. Wash the beads by resuspending them in 100 μL of Tris buffer. Reclaim beads, discard the buffer, and resuspend the beads in 50 μL of Tris buffer.
19. To amplify the Hi-C product, transfer 23 μL of the beads suspension (save the rest at 4 °C) to a PCR tube and add Illumina primers (1 μL of PCR primer 1 and 1 μL of PCR primer 2) plus 25 μL of KAPA HiFi master mix. Mix thoroughly and incubate the tube in a PCR machine under the following program (*see Note 7*)
 - Step 1: 98 °C for 30 s.
 - Step 2: 98 °C for 10 s.
 - Step 3: 65 °C for 30 s.
 - Step 4: 72 °C for 30 s.
 - Step 5: Go back to **step 2** and repeat 11 times.
 - Step 6: 72 °C for 5 min.
 - Step 7: 4 °C hold.
20. After the amplification is completed, bring the total library volume to 100 μL with H_2O .
21. Separate the beads on a magnet and transfer the solution to a fresh tube. Discard the beads.

22. Add 100 μL AMPure XP of beads to the solution ($1\times$ volume), mix by pipetting at least ten times, and incubate at room temperature for 5 min.
23. Separate the beads on a magnet and remove the clear solution.
24. Keeping the beads on the magnet, wash twice with 200 μL of 80% ethanol without mixing. Incubate the beads in ethanol for 30 s for each wash.
25. Remove ethanol completely. Leave the beads on the magnet for about 5 min to allow the remaining ethanol to evaporate. Do not over-dry the beads to prevent yield loss.
26. Resuspend the beads with 25 μL of Tris buffer and mix by pipetting; incubate at room temperature for 5 min. Separate the beads on a magnet, and transfer the solution to a new tube.
27. Quantify the library concentration using Qubit dsDNA High Sensitivity Assay and examine the library size distribution by a D1000 tape on the Tapestation or by Bio-analyzer (*see Note 7*).
28. Load the libraries on Illumina sequencing platform and identify significant chromatin interactions based on the sequencing results (*see Note 8*).

4 Notes

1. This Hi-C protocol aims to investigate most chromatin interactions in a cell population. As each chromatin interaction may be detected in every 1000 cells on average, low starting cell number may lead to low complexity Hi-C libraries. Therefore, researchers are encouraged to begin a Hi-C library with ten million cells.
2. Crosslinking is a critical step for a successful Hi-C library. Under-crosslinking may result in the disassembly of physical interactions, while over-crosslinking will impede the accessibility of restriction enzymes. Both the scenarios will lead to the underestimation of chromatin interactions in the final Hi-C results. Optimization of the fixation parameters such as the formaldehyde concentration (1–2%) and fixation time (5–10 min) is recommended for each cell type and biological status. Researchers may utilize the highest fixation parameter without interfering with digestion and ligation efficiency. The effect can be examined as described in Subheading 3.4.
3. This step is to fill in the restriction fragments overhangs and mark the DNA ends with biotin, make sure excessive biotin-dCTP is used.
4. After digestion, a smear should be seen on the agarose gel and this smear will focus on a certain size after the re-ligation step

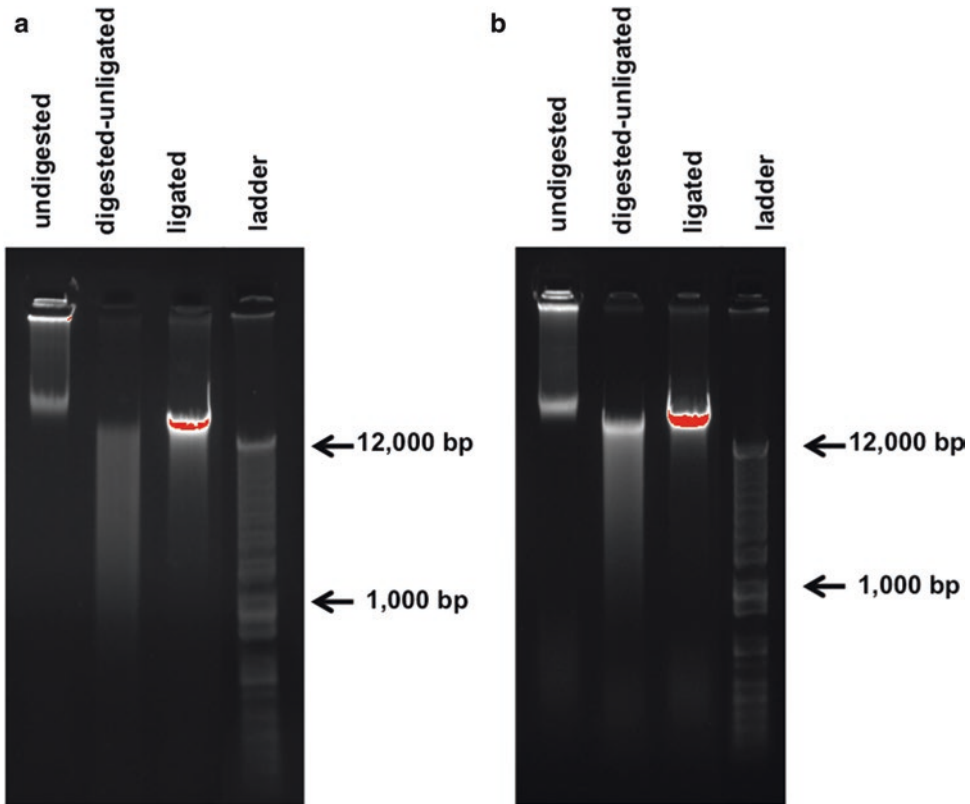


Fig. 1 Electrophoresis on undigested, digested-unligated, and ligated controls. Sample in (a) was fixed by 1% formaldehyde for 5 min and electrophoresis shows high efficiency of digestion whereas sample in (b) was fixed by 1% formaldehyde for 10 min and the digestion efficiency is lower

(usually on the average size of chromatin loops in a certain genome, but be aware that the relative position to the ladder may not represent the size of circular DNA generated during ligation). If the digestion efficiency is low, try to decrease the formaldehyde concentration or fixation time (*see* Fig. 1a, b).

5. After shearing, the DNA fragment size distribution should be around 200 bp (*see* Fig. 2a).
6. The supernatant in **step 10** contains shorter than 350 bp. In **step 12**, fragments in the range of 150–350 bp will be retained on the beads and the supernatant contains RNA and DNA fragments shorter than 150 bp. The Tapestation result after size selection is shown in Fig. 2b. Size selection could also be performed by running samples on 2% agarose gel and cutting the gel between 150 and 350 bp.
7. In order to obtain decent amount of library and avoid over-amplification, it is recommended to determine PCR cycle number prior to the first experiment. Perform 4, 8, 12 cycles

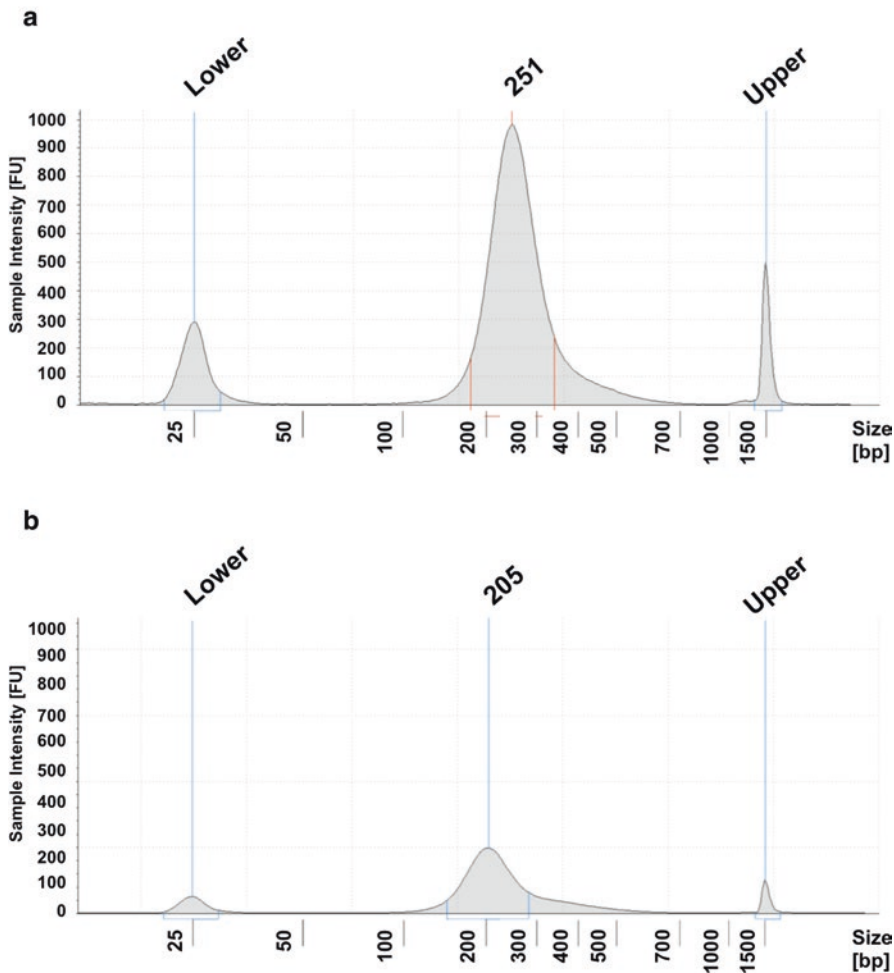


Fig. 2 Tape Station results on DNA before (a) and after (b) size selection

of amplification and examine library concentration and library size distribution. Examples of good and over-amplified library are shown here (*see* Fig. 3a, b).

8. Hi-C sequencing results consist of hundreds of millions reads on de novo ligation product. There are some published algorithms for Hi-C data processing and analysis, for example, Hi-Corrector [13] and HiCNorm [14] are focusing on HiC data normalization; HiC-Pro [15], diffHiC [16], HiFive [17], and HOMER [18] introduce whole pipelines of HiC data analysis; BACH [19] is specific for simulating 3D structures. However, we still hope to go through some basic procedure of Hi-C data here in order to make some fundamental concepts more clear. The workflow of Hi-C analysis begins with extracting valid interaction pairs and mapping them

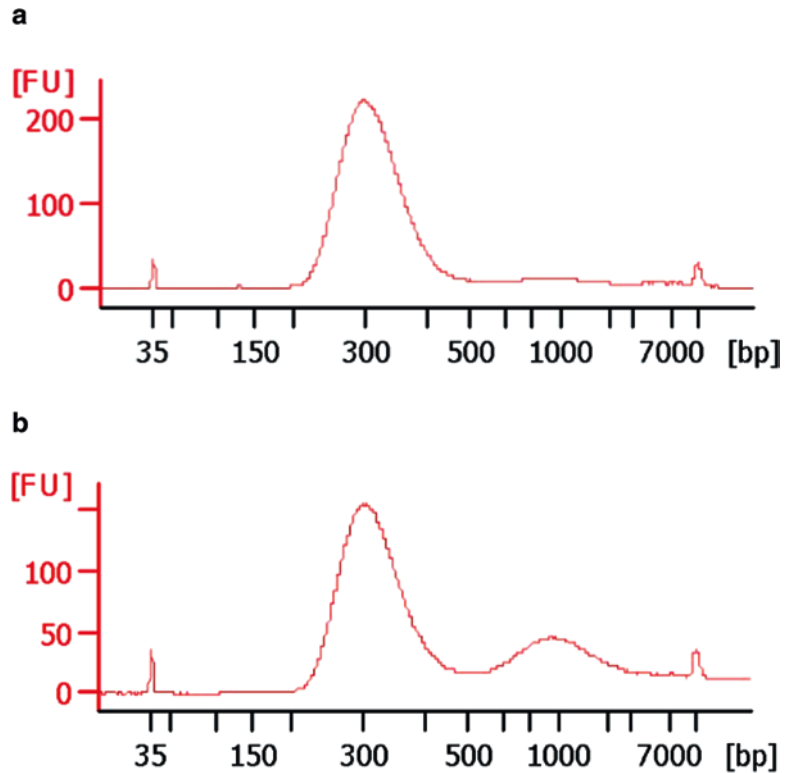


Fig. 3 Size distribution in non-over-amplified (a) and over-amplified (b) libraries

back to restriction sites over the whole genome, followed by generating a Hi-C contact matrix, identifying significant chromosome interactions and different chromatin structural domains [20, 21].

To map sequenced Hi-C reads to reference genome, any short read sequence alignment program can be used (e.g., Bowtie [22], SOAP [23], etc.), but before running alignment program, we recommend researchers to trim the reads at the de novo re-ligation sites (e.g., NheI sites generated from re-ligating Hind III digestion). Since reads spanning a de novo ligation site usually cannot be aligned to genome, trimming off the 3' part further than these sites could increase the unique mapping rate.

Valid interaction pairs are the ones uniquely aligned to different restriction fragments, facing toward the restriction site. Therefore, after alignment, reads are subject to filtering, in order to remove several experimental artifacts, which mainly consist of self-ligation, dangling ends, random breaking, and PCR amplification. Typically, self-circles comprise 0.5–5% of the molecules in a Hi-C library, and dangling ends may comprise 10–45%. PCR over-amplification may result in redundant reads that are aligned to the same genome posi-

tion at both the ends. In such cases, only one copy should be included. Typically, less than 5% of the library of valid pairs having redundant reads is acceptable; otherwise concern may be raised about PCR over-amplification.

Hi-C normalization is still a challenge for bio-informaticians. Currently, the most widely accepted Hi-C normalization method appears as iterative correction and eigenvector decomposition (ICE) [24]. ICE rescales the coefficients of the interaction matrix to make all rows and columns to have an equal sum. Even though it provides the most optimized treatment so far, limitations still exist. First, the diagonal and nearest neighbors are excluded while rescaling the coefficients, This is based on the assumption that any chromatin site interacts with a locus that is not its closest neighbor. Therefore, some true short-distance interactions may be eliminated; second, it cannot calculate confidence intervals for the interactions; third, it does not provide goodness-of-fit criterion, so whether ICE has over-cleaned a matrix or not cannot be determined.

After normalization, a contact matrix could be created which represents the overall patterns of chromosome interactions. In a successful Hi-C experiment, a strong diagonal of interactions between proximal genomic sites which shows overall decay over increasing distance and clear clusters of intra- and inter-chromosomal interactions could be observed.

Due to the huge volume and complexity of the human genome, most of Hi-C libraries on human cells are often under-sequenced. For instance, the human genome has about one million HindIII restriction sites (if HindIII is being used as endo-restricted enzyme in the experiment), so that 100 million valid pairs of reads could only provide 200 read ends at each HindIII restriction fragment on average. Theoretically, every HindIII fragment could physically interact with each of the other fragments thus only 0.0002 reads representing any two HindIII fragments contact could be obtained. This makes it difficult to determine whether a single observed interaction is a significant event or random noise. One current solution is to combine multiple biological replicates if they show high reproducibility (Pearson's correlation coefficient >0.9 at 1 Mb resolution) to compensate for the sequencing depth. Another idea is that in identifying significant interactions, reads across large regions are piled together in order to boost the total number of reads considered.

Acknowledgment

This is dedicated to the memory of Dr. Daniel R. Salomon. This work was supported by NIH grant 5U19 AI063603 and Mendez National Institute of Transplantation Foundation.

References

1. Tessarz P, Kouzarides T (2014) Histone core modifications regulating nucleosome structure and dynamics. *Nat Rev Mol Cell Biol* 15(11):703–708
2. Negre N et al (2011) A cis-regulatory map of the drosophila genome. *Nature* 471(7339):527–531
3. Li G et al (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148(1–2):84–98
4. Apostolou E et al (2013) Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. *Cell Stem Cell* 12(6):699–712
5. Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. *Science* 295(5558):1306–1311
6. Weiner BM, Kleckner N (1994) Chromosome pairing via multiple interstitial interactions before and during meiosis in yeast. *Cell* 77(7):977–991
7. Lieberman-Aiden E et al (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950):289–293
8. Dixon JR et al (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature* 518(7539):331–336
9. Dixon JR et al (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398):376–380
10. Rao SS et al (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159(7):1665–1680
11. Jin F et al (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 503(7475):290–294
12. Duan Z et al (2012) A genome-wide 3C-method for characterizing the three-dimensional architectures of genomes. *Methods* 58(3):277–288
13. Li W, Gong K, Li Q, Alber F, Zhou XJ (2015) Hi-corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics* 31(6):960–962
14. Hu M et al (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* 28(23):3131–3133
15. Servant N et al (2015) HiC-pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* 16:259
16. Lun AT, Smyth GK (2015) diffHic: a bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* 16:258
17. Sauria ME, Phillips-Cremens JE, Corces VG, Taylor J (2015) HiFive: a tool suite for easy and efficient HiC and 5C data analysis. *Genome Biol* 16:237
18. Heinz S et al (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38(4):576–589
19. Hu M et al (2013) Bayesian inference of spatial organizations of chromosomes. *PLoS Comput Biol* 9(1):e1002893
20. Lajoie BR, Dekker J, Kaplan N (2015) The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* 72:65–75
21. Dekker J, Marti-Renom MA, Mirny LA (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* 14(6):390–403
22. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25
23. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics* 24(5):713–714
24. Imaekave M et al (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* 9(10):999–1003

Primer Extension, Capture, and On-Bead cDNA Ligation: An Efficient RNAseq Library Prep Method for Determining Reverse Transcription Termination Sites

Phillip Ordoukhanian, Jessica Nichols, and Steven R. Head

Abstract

In this chapter, we describe a method for making Illumina-compatible sequencing libraries from RNA. This protocol can be used for standard RNAseq analysis for detecting differentially expressed genes. In addition, this protocol is ideally suited for adapting to RIPseq, 5'-RACE, RNA structural probing, nascent RNA sequencing, and other protocols where polymerase termination sites need to be profiled. The utilization of solid-phase bead chemistries facilitates simple workflow and efficient library yields.

Key words RNA-seq, Library preparation, Ribo-seq, Gene expression, Transcriptome

1 Introduction

Numerous methods have been developed for the creation of sequencing libraries from RNA samples [1–4]. The method here was developed to take advantage of the efficiencies made possible by bead-based purification and heterogenous phase enzymatic reactions. In this method, adapters are attached to the library using adapter-primer extension, bead capture, and on bead adapter-ligation. Similarly, published methods have used in solution cDNA adapter ligation to identify RNA/protein crosslinking sites [5] and others have used adapter-primer extension, capture, and a second on bead adapter-primer extension to generate RNAseq libraries [6]. The method here differs in that cDNA adapter-ligation is used instead of random priming and the ligation reaction occurs on-bead, greatly simplifying the protocol. The cDNA ligation step allows for the determination of the reverse transcription termination site. This is useful for several applications, such as mapping RNA modifications and protein–RNA cross-links [7, 8], structural probing of RNAs [9], nascent RNA [10, 11], or rapid amplification of cDNA 5' ends [12, 13].

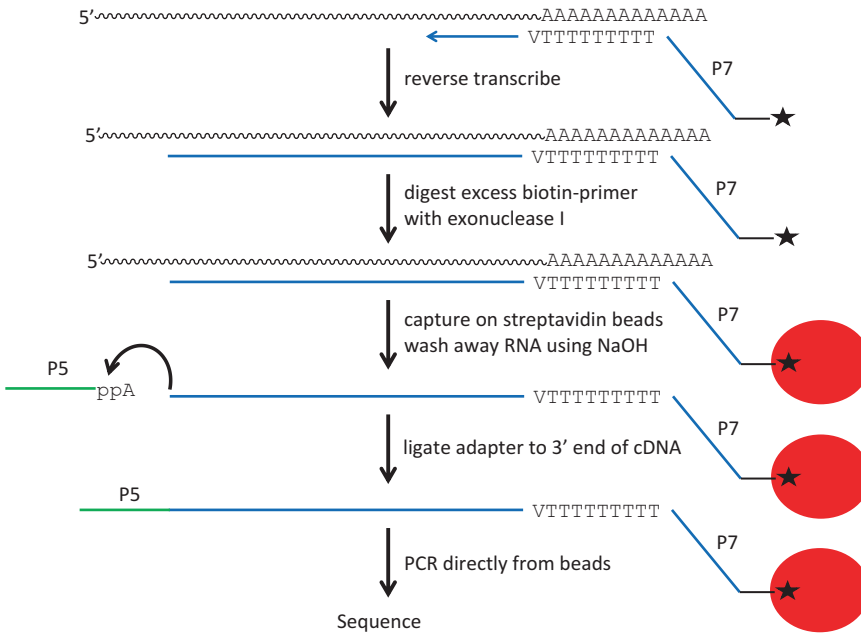


Fig. 1 Library prep workflow

We have routinely used this library prep method with as little as 10 ng of RNA input. For total RNA samples requiring ribo-depletion or oligodT selection, we recommend starting with at least 100 ng. If working with total RNA samples, one should either ribo-deplete or oligo-dT select the sample before beginning the protocol. In the protocol, the RNA is then lightly fragmented with divalent ions and heat (*see* Fig. 1). Next, a 5'-biotinylated adapter-primer, terminated in either a random octamer or oligo-dT, is used to perform cDNA synthesis on the RNA library. The excess biotinylated primer is digested with exonuclease I. Solid-phase Reverse Immobilization (SPRI) beads are used to purify the cDNA/RNA duplex from the reaction products and buffer components. The cDNA/RNA duplex is then bound to streptavidin-magnetic beads and washed with 0.1 N sodium hydroxide solution to wash away the RNA strand. After washes with neutralizing buffer, a second adapter is ligated directly to the 3'-end of the cDNA bound to streptavidin-magnetic beads using a thermostable ligase and an adenylated DNA adapter. Following this reaction, the beads are washed several times with buffer and added directly to a PCR reaction to generate functional Illumina compatible sequencing libraries.

2 Materials

2.1 Equipment and Supplies

1. Heated-lid thermocycler.
2. Magnetic stand for microfuge tubes.
3. 2100 Bioanalyzer, Tapestation (Agilent Technologies, Santa Clara, CA, USA) or other equivalent systems, e.g., Fragment Analyzer (Advanced Analytical, Ankeny, IA, USA) (*see Note 1*).
4. Qubit[®] 2.0 Fluorometer, Assay Tubes, and dsDNA HS Assay Kit (ThermoFisher Scientific, Waltham, MA, USA).
5. Egel[®] apparatus and 2% Egel EX gels (ThermoFisher Scientific) or other submerged agarose gel running apparatus (*see Note 2*).
6. Blue light transluminator (*see Note 3*).
7. DNA Clean & Concentrator[™]-25—DNA, includes Agarose Dissolving Buffer (ADB) and DNA Wash Buffer (Zymo Research, Irvine, CA, USA) (*see Note 4*).

2.2 Reagents and Consumables

1. ILMN-primer-adapter dT: /5Biosg/GAGTTCCTTGGCA C C C G A G A A T T C C A T V (*see Note 5*).
2. ILMN-primer-adapter random: /5Biosg/GAGTTCCTTGG CACCCGAGAATTCCA(N:25,252,525)(N)(N)(N) (N)(N) (N)(N) (*see Note 5*).
3. ILMN-ligation-adapter: /5rApp/(N:25,252,525)(N)(N) GATCGTCGGACTGTAGAACTCTGAACGTGT/3ddC/ (*see Note 5*).
4. PCR primer 1: AATGATACGGCGACCACCGAGATCTAC ACGTTCAGAGTTCTACAGTCCGACGATC.
5. PCR primer 2 with sample barcode (X = Illumina barcodes): CAAGCAGAAGACGGCATAACGAGATXXXXXXXXGT GACTGGAGTTCCTTGGCACCCGAGAATTCCA.
6. Low DNA binding microfuge tubes.
7. Thin-walled PCR microfuge tubes.
8. 2% Agarose Egel EX (ThermoFisher Scientific).
9. RNaseOUT[™] (ThermoFisher Scientific).
10. Dynabeads Oligo(dT)₂₅ beads (ThermoFisher Scientific).
11. 5 M Lithium Chloride.
12. 1 M Tris-HCl pH 7.5.
13. 0.5 M EDTA.
14. High-Salt (HS) buffer: 2 M NaCl, 10 mM Tris-HCl (pH 8), 1 mM EDTA.
15. RNA Binding buffer: 20 mM Tris-HCl (pH 7.5), 1 M LiCl, 2 mM EDTA.

16. Wash Buffer 1: 10 mM Tris-HCl (pH 7.5), 150 mM LiCl, 1 mM EDTA.
17. Wash Buffer 2: 10 mM Tris-HCl (pH 7.5), 20 mM KCl.
18. M280 Streptavidin Dynabeads (ThermoFisher Scientific).
19. 10 mM dNTP solution.
20. Superscript III (ThermoFisher Scientific).
21. Thermostable App DNA/RNA Ligase (New England Biolabs, Ipswich, MA, USA).
22. 50 mM MnCl₂.
23. 10× NEB1 buffer.
24. Exonuclease I (New England Biolabs).
25. KAPA HiFi HotStart DNA polymerase (KAPA Biosystems, Wilmington, MA, USA).
26. AmpureXP beads (Beckman Coulter, Carlsbad, CA, USA).
27. 2 N sodium hydroxide solution (dilute for 0.1 N NaOH solution).
28. 20× TE (dilute for 1× TE buffer).
29. 80% ethanol.

3 Methods

3.1 Poly A Selection of mRNA (Optional)

1. Vigorously resuspend oligo-dT beads, pipette 20 μL into 0.5 mL Lo-Bind microfuge tube and place in magnetic stand, once the beads have cleared the solution, remove the supernatant. Wash the beads 2 times with 100 μL RNA Binding Buffer. Then, resuspend in 30 μL RNA binding buffer.
2. Add at least 100 ng total RNA in a volume of 10 μL to the tube containing the 30 μL suspension of magnetic beads above and pipette mix.
3. Place the tube in a thermocycler and incubate at 65 °C for 5 min, then lower to 4 °C. Remove the tube from thermocycler and incubate at room temperature for 7 min, to further allow the mRNA to bind to the beads. Place the tube in a magnetic stand, remove and discard the supernatant.
4. Using pipette mixing—wash the beads two times with 100 μL Wash Buffer 1.
5. Add 20 μL of 10 mM Tris pH 7.5, pipette mix. Heat to 80 °C for 2 min, then hold at 25 °C.
6. Add 30 μL RNA Binding Buffer, pipette mix. Incubate at room temperature for 7 min to allow mRNA to re-bind to beads. Transfer the tube to magnetic rack and remove and discard the supernatant.

7. Pipette mix-wash beads 1 time with 100 μ L Wash Buffer 1 and 1 time with 100 μ L Wash Buffer 2. Remove and discard each wash.
8. Elute RNA—Add 10 μ L of 10 mM Tris pH 7.5, pipette mix. Heat to 80 °C for 2 min, then immediately transfer to a magnet and remove the supernatant containing the mRNA and place in a fresh thin-walled PCR tube for subsequent reverse transcription below (*see* **Note 6**).

3.2 Reverse Transcription

1. To the 10 μ L of eluted mRNA from Subheading 3.1 add 1 μ L ILMN-primer-adaptor dT (20 μ M) and 4 μ L of 5 \times First Strand Buffer.
2. Incubate at 95 °C, 3 min, then place on ice for at least 3 min.
3. Prepare a *master mix* for the total number of samples being prepped in a separate 0.5 mL microfuge tube, containing the following reagents (due to pipetting inaccuracies, an additional 10% in volume per sample prep per reagent is typically used): 2 μ L DTT (0.1 M), 1 μ L RNaseOUT™, 1 μ L dNTPs (10 mM), 1 μ L Superscript III. 5 μ L of this master mix is then added to each sample from the previous step. The final volume will be 20 μ L.
4. Incubate at 25 °C for 15 min, then increase to 42 °C for 40 min, and finally to 75 °C for 10 min.
5. Digest excess primer: Add 2 μ L Exonuclease I and incubate at 37 °C for an additional 30 min.
6. Isolate cDNA/RNA material: Add 5 μ L nuclease-free water and 45 μ L AmpureXP beads to the tube containing the reaction from above. Pipette mix up and down ten times to mix. Allow cDNA/RNA to bind to the beads for 10 min, place the tube on a magnetic stand, and remove the supernatant. Wash the beads 2 times with 80% ethanol while on the magnetic stand (100 μ L each). Allow the beads to air-dry at ambient temperature for 10 min. Remove the tube from the magnetic stand and add 25 μ L 10 mM Tris-HCl (pH 8.0) to the beads, pipette mix, and let it sit at ambient temperature for 2 min. Return the tube to the magnetic stand and retrieve 24 μ L of cDNA/RNA material.

3.3 Streptavidin Bead Capture

1. Prepare M280-Streptavidin Dynabeads: Add 25 μ L M280 Streptavidin Dynabeads to a fresh tube. Place in a magnetic stand and remove the supernatant. Wash the beads 1 time with 200 μ L HS buffer. Remove the tube from the magnetic stand and resuspend Streptavidin Beads in 25 μ L HS buffer.
2. Combine: 25 μ L eluted cDNA/RNA material with 25 μ L washed Streptavidin Beads and incubate at ambient temperature for 30 min (pipette mix after 15 min).

3. Place the tube in a magnetic stand and remove the supernatant, then wash Streptavidin Bead-bound cDNA/RNA one time with a 0.1 N NaOH solution (100 μ L), followed by two washes with 1 \times TE buffer (100 μ L each). Remove the last wash and discard, then leave the streptavidin beads with bound cDNA in the tube, and place the tube on ice while preparing the next reaction mixture.

3.4 A-Adapter Ligation

1. Prepare a *master mix* for the total number of samples being prepped in a separate 0.5 mL microfuge tube, containing the following reagents (due to pipetting inaccuracies, an additional 10% in volume per sample prep per reagent is typically used): 2 μ L 10 \times NEB1 buffer, 2 μ L 50 mM MnCl₂, 2 μ L ILMN-ligation-adapter (50 μ M), 12 μ L nuclease-free water, 2 μ L ThermoStable App DNA/RNA Ligase.
2. To the streptavidin beads with the bound cDNA from **Step 3**, Subheading 3.3 transfer 20 μ L of the reaction cocktail prepared above and pipette up and down until mixed well. Incubate at 60 °C for 1 h.
3. Wash the streptavidin bead bound cDNA: 3 washes with 1 \times TE buffer (100 μ L each), and resuspend the streptavidin beads in 36 μ L nuclease-free water, creating a slurry, then transfer the slurry to a fresh thin-walled PCR tube.

3.5 PCR and DNA Purification

1. Prepare a *master mix* for the total number of samples being prepped in a separate 0.5 mL microfuge tube, containing the following reagents (due to pipetting inaccuracies, an additional 10% in volume per sample prep per reagent is typically used): 2 μ L PCR primer 1 (25 μ M), 2 μ L PCR primer 2 with sample barcode (25 μ M), and 40 μ L 2 X KAPA HiFi mix and pipette mix up and down. Transfer 44 μ L of the reaction cocktail to the thin-walled PCR tube containing the cDNA bound streptavidin beads. The final volume will be 80 μ L.
2. Perform thermocycling in a PCR instrument: 1 hold—95 °C, 3 min; 15 cycles—98 °C for 20 s, 60 °C for 30 s, 72 °C, 30 s; 1 hold—72 °C, 2 min.
3. Following the PCR, add 80 μ L AmpureXP beads (1 \times strength) directly to the PCR tube and pipette mix. Let it stand for 10 min to precipitate the DNA onto the beads. Place the PCR tube in a magnetic stand and remove the supernatant. With the tube still in the magnetic stand wash the beads two times with 80% ethanol solution (200 μ L each). Allow the beads to air-dry with the lid of the tube open at ambient temperature for 10 min. Remove the tube from magnetic stand and add 25 μ L 10 mM Tris-HCl (pH 8.0), pipette mix. Let it stand for 2 min, return the tube to the magnetic stand, and allow the beads to clear the solution, transfer 24 μ L to a fresh microfuge tube (avoiding removal of any beads).

4. PCR DNA products should be analyzed and quantitated on a DNA fragment analyzer or by agarose gel and purification (*see item 3*, Subheading **2.1** and **Note 1**). Perform a smear analysis quantitation on region between 250–600 bp.
5. Pool all samples equimolar based on quantitation. Load the pooled sample onto one lane of a 2% Agarose Egel EX (*see Note 2*). Be sure not to overload the lane, for an Egel EX keep the loading between 50 and 100 ng of pooled DNA material. Electrophorese until markers are at desired distance, remove the cassette and break it open.
6. Visualize the smear of DNA products on a blue light transilluminator and using a razor blade, excise the piece of agarose gel into a microfuge tube. Use the excising blade to cut up the gel slice into smaller pieces (this will decrease the time it takes to dissolve the gel slice). Weigh the gel slice in the microfuge tube. Based on the weight, add four volumes of Agarose Dissolving Buffer (ADB) (e.g., if the gel weighs 100 mg, add 400 μ L). Place the tube in a 37 °C heat block and incubate for 10 min (remove the tube and flick-mix it every few min). Load dissolved agarose gel and buffer onto a Zymo-25 column and centrifuge at recommended speed. Wash the column two times with 200 μ L DNA Wash Buffer. Centrifuge for 30 s on the first wash and for 1.5 min on the second to ensure all wash buffer is removed before adding the elution buffer. Elute the DNA products with 30 μ L 10 mM Tris pH 8 into a fresh microfuge tube. Quantitate products by Qubit analysis and run a Fragment Analyzer on them to confirm product sizes.
7. The DNA products are ready to be taken into the standard Illumina sequencing protocol and sequenced.

4 Notes

1. Any system that analyzes fragment length can be substituted. However, just an agarose gel will suffice, but each sample must be purified separately then, quantitated separately and then pooled together based on the quantitation. In contrast, smear analysis from a fragment analyzer can be used prior to purification to quantitate the smear of desired fragment sizes and based on that quantitation pool the samples, and then purify the one pooled sample.
2. Standard submerged gel apparatuses can be used, but care should be taken to avoid well-to-well contamination, such as from the DNA ladders or other libraries or other samples that have the same barcodes. Egels do not use low-melting agarose, but the protocol followed here will dissolve the gels. In our

experience, low-melt agarose tends to smear the products more and therefore complicate the purification. Standard high-melt agarose can be used with this protocol.

3. Blue light transluminator must be used so DNA is not damaged. A standard transluminator is not recommended.
4. Other DNA clean-up columns may be substituted; however, if the agarose gel is not allowed enough time to dissolve completely, tiny bits can clog some columns. In that case, try adding a little more buffer to the column, wait a short while, and try centrifuging again. In our experience, we generally use the Zymo-25 columns, the Zymo-5 column tends to clog more often than the Zymo-25.
5. Where N's are hand-mixed random nucleotides (hand-mixed random nucleotides have a more even distribution of the nucleotides and were used to avoid potential bias problems), V is every other DNA base except T, /5Biosg/ is a 5'-biotin derivative, /5rApp/ is a 5'-adenylation, and /3ddC/ is 3' dideoxy-cytidine derivative.
6. The transfer to a fresh thin-walled PCR tube must be performed immediately after heating 80 °C for 2 min. Failure to transfer in a timely manner will allow the mRNA to re-bind to the magnetic beads causing a drop in yield. If doing multiple samples at one time, no more than 2 samples should be done at a given time.

References

1. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P (2014) Library construction for next-generation sequencing: overviews and challenges. *BioTechniques* 56(2):61–77
2. Podnar J, Deiderick H, Huerta G, Hunicke-Smith S (2014) Next-generation sequencing RNA-Seq library construction. *Curr Protoc Mol Biol* 106:4.21.1–4.21.19
3. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A (2009) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* 7(9):709–715
4. Routh A, Head SR, Ordoukhanian P, Johnson JE (2015) Fragmentation-free next-generation sequencing via click-ligation of adaptors to stochastically terminated 3'-azido cDNAs. *J Mol Biol* 427(16):2610–2616. <https://doi.org/10.1016/j.jmb.2015.06.011>
5. Kieplinski LJ, Boyd M, Sandelin A, Vinther J (2013) Detection of reverse transcriptase termination sites using cDNA ligation and massive parallel sequencing. *Methods Mol Biol* 1038:213–231. https://doi.org/10.1007/978-1-62703-514-9_13
6. Li H, Lovcib MT, Kwona YS, Rosenfeld MG, Fua XD, Yeo GW (2008) Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proc Natl Acad Sci U S A* 105(51):20179–20184
7. König J, Zarnack K, Luscombe NM, Ule J (2012) Protein–RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet* 13:77–83. <https://doi.org/10.1038/nrg3141>
8. Linder B, Grozhik AV, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR (2015) Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome.

- Nat Methods 12:767–772. <https://doi.org/10.1038/nmeth.3453>
9. Loughrey D, Watters KE, Settle AH, Lucks JB (2014) SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Res* 42(21):e165
 10. Khodor YL, Rodriguez J, Abruzzi KC, Tang CHA, Marr MT II, Rosbash M (2015) Nascent-seq indicates widespread co-transcriptional pre-mRNA splicing in drosophila. *Genes Dev* 25:2502–2512
 11. Chen F, Gao X, Shilatifard A (2015) Stably paused genes revealed through inhibition of transcription initiation by the TFIID inhibitor triptolide. *Genes Dev* 29(1):39–47. <https://doi.org/10.1101/gad.246173.114>
 12. Frohman MA (1993) Rapid amplification of complementary DNA ends for generation of full-length complementary DNAs: thermal RACE. *Methods Enzymol* 218:340–356
 13. Loh E (1991) Anchored PCR: amplification with single-sided specificity. *Methods* 2:11–19

INDEX

A

Activation-induced chromatin structural rearrangement.....vii
 Adapter-dimersvii, 41, 147, 150, 151, 156, 158, 160, 161
 Alkyne 71, 74, 77
 Alkyne-labeled DNA oligos74
 Amplicon-sequencing.....45
 Ampliseq53, 123
 Antigen-driven immune responses.....217
 Archaea.....99
 Array comparative genomic hybridization (arrayCGH).....27
 Autoimmunity.....217, 218
 Azido..... 71, 74, 76, 77, 80, 82, 83
 3'-Azido blocked cDNA fragments72
 3'-Azido-nucleotides.....72

B

Bacterial type II clustered regularly interspaced short palindrome repeats (CRISPR)203
 Baits..... vii, 44, 46–49
 BAM file 50, 121–124, 127, 128, 131, 182
 B cells2, 7, 9, 10, 191
 Bead-based purification..... 149, 160, 253
 Bioinformatics viii, 34, 50, 93, 94, 99, 114
 Bio-orthogonal chemistry.....71
 Biotin-streptavidin44
 Bisulfite vii, 87, 93
 Bisulfite sequencing on single cells (scBS-Seq)88
 BLAST.....99, 108, 115, 137, 139
 Blastomere.....27
 Blood..... 170, 176, 188, 195–199, 205, 211, 226

C

5' Capping.....19
 Cas9.....203, 204
 CD4+191
 CD8+ T cells191
 cDNA synthesis..... 72, 74, 77, 146, 149, 152, 153, 219, 237, 254
 Cell fixation.....247
 Cell lysis90, 98, 100, 102–104, 242

CellCODE package188, 189, 193, 195, 197
 Cell type-specific deconvolution.....175
 Cellular heterogeneity 176, 200
 CheckM 99, 108
 Chimerasviii, 71–74, 79, 82, 98
 Chromatinvii, 20, 22, 25, 239, 247, 248, 250
 Chromosomal copy number alterations (CNAs)..... vii, 28, 34, 37, 39
 Chromosomal interaction.....239–251
 Chromosome conformation capture (3C)239
 CircLigase6
 Cis-regulatory elements.....88
 CleanTag vii, 145–161
 Click-chemistry.....viii, 71–73, 79, 82
 Click-ligate..... 72, 78, 82, 83
 ClickSeq viii, 71–84
 Complex microbial communities.....171
 Conda.....178, 189
 Contigs..... 98, 99, 108, 116, 117, 120–123
 Copper-catalyzed Azide-Alkyne Huisgen Cycloaddition (CuAAC) 71, 72, 74
 CpG methylation88
 CRISPR-Cas9.....203–205, 208–210
 CRISPR genomic screens.....212, 214
 Cultivation-independent molecular tools97
 Cyclohexamide (CHX).....3

D

DeconSeq 99, 108
 Decontamination..... 4, 17, 90, 99, 100, 102, 103, 108
 Defective-interfering viral RNAs71
 Delphinidae116
de novo ligation 240, 248
 Dithiothreitol (DTT).....5
 DNA extraction..... 55, 163–169, 172, 205, 210, 241, 243–245
 DNA methylation87–94
 Droplet-based microfluidic.....218

E

Elongation 2, 19, 48, 92
 Embryo..... 27, 28, 205
 Encyclopedia of DNA elements (ENCODE).....87, 180

Endo-restricted enzyme digestion..... 239, 240
 Enhancers..... 33, 88
 Environmental samples 97, 102, 103, 109
 Enzymes..... 30–33, 39, 46–48, 54, 61, 67, 74, 81,
 90, 91, 94, 146, 149, 151, 152, 159, 171, 212–215, 219,
 220, 242, 247, 251
 Epigenetic 87, 88
 Eukaryotic cells 71, 170, 239

F

Fastq..... 35, 44, 50, 115, 119–121, 125–127,
 179–183, 185, 186, 188, 201
 FastQC..... 93, 177, 179, 180, 201
 Feces..... 163, 168, 169
 Flock House virus..... 74, 81
 Fluidigm..... 53, 218–220, 227, 237, 238
 Fluidigm Biomark HD..... 218
 Fluorescence-activated cell sorting (FACS)..... 217
 Functional variants 113

G

Gel excision 53
 Gel-free 146
 Gene expression..... 2, 87, 188, 189, 193, 198, 200,
 217, 237
 Gene-specific primers (GSPs)..... 54–56, 60, 61, 64, 66,
 67, 69
 Genetic variant vii, 54, 66
 Genome-Scale CRISPR Knock
 Out (GeCKO)..... vii, 204, 207–210, 213
 Genotypes 113, 114, 124, 125, 128, 129, 133, 137
 Github..... 178
 Glimnet..... 189, 190, 193–195, 198
 Global run-on sequencing (GRO-seq)..... 19
 Gut microbiome 165, 168, 169

H

Haloplex 54
 Hi-C..... vii, 239,
 249–251
 Highest fidelity PCR polymerases 53
 Hi-Plex..... vii, 53–69
 Hybridization 44, 46–48

I

Illumina NextSeq 500 29, 39, 207, 214
 Illumina sequencing libraries 28
In vitro..... 7, 20, 71
 Isoform 177, 180, 182

K

K-mer frequencies 99

L

Lambda Phage 44
 Library preparation..... viii, 4–6, 9–16, 20, 23, 28–30,
 32–34, 39, 45, 46, 49, 50, 98, 99, 101, 102, 106, 107, 147,
 150, 151, 156, 158, 160, 161, 171, 206, 207, 211–214,
 229, 255, 260
 Ligation..... 5, 9, 10, 12, 33, 40, 44, 45, 49, 71–74,
 77–83, 98, 146, 149, 151, 152, 159, 240–243, 246, 248,
 253–260
Limma 189–193, 198
 Linux 114–117, 119, 121, 139, 185

M

Mammalian cells 87, 93, 203, 207
 Massive parallel sequencing (MPS)..... 27, 28
 Metabolic reconstruction..... 99
 Metagenomics 97
 Microarray 175, 203
 Microbes..... 97, 101, 164, 167
 Microbial genomes 97
 Microbiome..... 123, 163–168
 Microfluidics viii, 98, 217
 Micropipetting 98
 MicroRNA (miRNA)..... 145, 148, 154, 159
 MinION..... vii, 45, 50
 Mobile elements 45
 Monocytes 191
 mRNA..... 1–3, 7, 9, 19, 74, 81, 256, 257, 260
 Multiple annealing and looping based amplification
 cycles (MALBAC) 28
 Multiple displacement amplification (MDA)..... 28,
 98–100, 102, 103, 105, 106, 108, 109
 Multiplex PCR..... 53
 Mutation screening..... 53

N

Nanopores vii, 45, 50
 Nascent-seq 19–25
 Native elongating transcript sequencing (NET-seq) 20,
 204
 Neutrophils..... 191
 Next-generation sequencing (NGS)..... vii, viii, 20, 24,
 72, 74, 87, 106, 114, 121, 146, 147, 150, 151, 156,
 158–161, 203–215, 218
 NIH Roadmap Epigenomics..... 87
 NK cells..... 191
 Noncoding RNAs..... 19
 Non-homologous end-joining (NHEJ)..... 203
 Non-model organisms viii

O

Off-target priming 54, 67

Optical tweezers98
 Oral microbiome 164, 166, 167
Orcinus orca 116, 118, 119
 Oxford nanopore vii, 45, 50

P

Paired-end sequencing..... 54, 66
 Parental imprinting87
 PCR barcoding..... 44, 49, 50
 Penalized regression..... 190, 193–196
 Peripheral blood241
 Phusion polymerase..... 5, 6, 15, 214
 Phylogenetic diversity.....97
 Phylogenetic screening 98, 99, 101, 102, 105, 106, 109
 Phylogenetics.....97–99, 102, 109, 113
 Picoplex/SurePlex.....28, 29
 PiwiRNA (piRNA) 145, 154
 Pol II20
 Polyacrylamide5, 6, 10, 13, 15
 Polyadenylation19
 Polymerase recruitment19
 Polysome vii, 1–3, 9, 16
 Polysome profiling.....1–17
 Population genetics..... 113, 129
 Precision nuclear run-on sequencing (PRO-seq)19
 Preimplantation genetic diagnosis (PGD).....27–41
 ProDeGe 99, 108
 Prokaryotic adaptive immune system203
 Python.....50, 114, 115, 124, 178

R

5'-RACE, RNA structural probing253
 RADseq.....114
 Random amplification..... 166, 171
 Recombination viii, 74
 Reference-guided assembly113
 Repetitive elements 87, 127
 Ribo-seq1, 253
 Ribosome..... vii, 3, 4, 8, 10, 13, 17
 footprint abundance, 2
 profiling, 3, 4, 8, 10, 13, 17
 RNA interference204
 RNA sequencing viii, 9, 20, 21, 175–178, 180, 183, 185, 186,
 188, 191, 195, 197, 201
 RNase inhibitor3, 4, 9, 21, 22, 149, 151, 152, 159
 RNA-seq by expectation-maximization (RSEM) ... 177, 180,
 182, 185, 201
 R statistical computing environment 176, 177

S

Sample preparation..... 3, 24, 25, 99–102, 106, 109,
 226, 227
 Sample preservation98, 102
 Sanger sequencing 98, 106, 114

Sequence error28
 Sequence-screening53
 Short guide RNAs (sgRNAs).....203, 204, 207, 208, 211
 Short-hairpin RNAs (shRNAs) 204, 208
 Shotgun metagenomics97
 Single amplified genome (SAG)98, 99, 101, 106–108
 Single cell amplification104
 Single-cell analysis.....vii
 Single-cell isolation98
 Single nucleotide polymorphism (SNP)..... 113–140
 Size-selection.....30, 33, 41, 53, 54, 62–64, 79, 80, 146,
 148, 154, 157, 160, 229, 241, 243–245, 248, 250
 Small non-coding RNAs145
 Small RNA-sequencing (sRNA-Seq) 145, 148
 Snakemake 177–182, 187, 201
 Soil 101, 163, 165, 170
 Spliced transcripts alignment
 to a reference (STAR)..... 177, 180, 181, 185, 188, 201
 16S rRNA gene97–99, 101, 102, 105, 169
 ssDNA.....74
 16S/18S ribosomal RNA gene sequences 165, 166, 171
 Strain-promoted Azide-Alkyne Cycloaddition
 (SPAAC) 71, 72

T

Tagmentation 102, 106
 Targeted sequencing..... vii, 43–50, 53
 Target-enrichment46
 T cell receptor (TCR)..... 217–237
 T cells 239, 249, 250
 TCR sequencing.....228–231, 233, 235, 237
 Termination.....viii, 19, 81, 255, 260
 Three-dimensional DNA fluorescence
 in situ hybridization (3D-FISH).....239
 T4 polynucleotide kinase.....5
 Transcription viii, 5, 6, 12–15, 19, 20, 74–77,
 80, 81, 146, 149, 171, 218, 219, 227, 237, 239, 255, 260
 Transcriptional initiation19
 Translational regulation of gene expression1
 Transfer RNA (tRNA)30, 33, 41, 145, 154
 Translation.....1–3, 8
 Translatome.....3, 4, 8, 10, 13, 17
 Transposase 98, 106
 T4 RNA ligase 2 5, 12, 149
 Trophectoderm27, 28
 TruSeq Amplicon53

U

Unix..... 114, 115
 Unnatural triazole-linked backbones.....71
 Uracil88
 Urea 5, 20, 22
 3'UTR..... 2, 3, 74, 81
 5'UTR.....2, 3

V

Vaccines218
Vaginal microbiome..... 165, 168
Viral purification 165, 166, 170, 171
Viral-like particles (VLPs) 165, 170

W

Whole genome amplification (WGA)vii, 28–32, 40, 98,
100, 103–106

Whole genome association studies113
Whole-genome sequencing27,
44, 114

X

X-chromosome inactivation87

Y

Y RNA145