

Methods in
Molecular Biology 1706

Springer Protocols

Johanna K. DiStefano *Editor*

Disease Gene Identification

Methods and Protocols

Second Edition

 Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor

John M. Walker

**School of Life and Medical Sciences,
University of Hertfordshire, Hatfield,
Hertfordshire AL10 9AB, UK**

For further volumes:

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

Disease Gene Identification

Methods and Protocols

Second Edition

Edited by

Johanna K. DiStefano

Translational Genomics Research Institute, Phoenix, AZ, USA

 **Humana Press**

Editor

Johanna K. DiStefano
Translational Genomics Research Institute
Phoenix, AZ, USA

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-7470-2 ISBN 978-1-4939-7471-9 (eBook)
<https://doi.org/10.1007/978-1-4939-7471-9>

Library of Congress Control Number: 2017964097

© Springer Science+Business Media, LLC 2018

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

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

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

Printed on acid-free paper

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

Preface

Completion of the Human Genome Project (HGP) not only yielded a greater appreciation of the role of DNA in shaping species development and evolution, biology, and disease susceptibility, but also helped spawn technological advances that have revolutionized the field of human genetics. Perhaps the most significant impact of this endeavor has been on the manner in which researchers investigate the causes of complex human diseases. Efforts to characterize the genetic variation in the human genome have led directly to the development and application of a diverse range of technological and bioinformatics approaches to identify the roles of both rare and common alleles in complex disease. Such strategies range from genome-wide association studies to whole genome sequencing, and everything in between.

Over three thousand genetic mutations have been identified that contribute to the pathogenesis of highly penetrant human diseases. Efforts to uncover the genomic basis of rare conditions have been successful due to the less complicated genetics of monogenic diseases compared to complex disorders. In rare conditions, a single mutation, inherited in a simple manner between generations in affected families, is typically sufficient to cause disease. In contrast, complex diseases, such as diabetes, heart disease, neurological disorders, and most kinds of cancer, result from a complicated interaction of multiple genetic and environmental determinants, none of which are amenable to identification and characterization using the traditional approaches to monogenic disease gene discovery. Recent efforts to characterize genetic variation in the human genome, coupled with the rapidly developing field of genomics, have led directly to the development of new and innovative approaches to the identification of genes contributing to complex human diseases. This volume was prepared to present molecular methodologies used in the process of identifying a disease gene, from the initial stages of study design to locus identification and target characterization. The need for such a book derives from the intellectual revolution in biomedical science and the realization that the molecular determinants of human diseases are rapidly becoming identifiable through well-planned, technologically advanced approaches.

While descriptions of the technical procedures described here are available in the literature, there is generally a dearth of practical detail in these publications, particularly in terms of modifications developed from personal experience and discussions of optimal study design or potential problems that may be encountered throughout the protocol, as well as ways to avoid them. The structure of this volume is unique in that it aims to address these deficiencies.

This text is written at a level accessible to graduate students, postdoctoral researchers, and bench scientists in the fields of molecular genes and molecular biology. The primary aim of this volume is to present detailed laboratory procedures in an easy to follow format that can be carried out with success by competent investigators lacking previous exposure to a specific research method. The book's main focus is on the application of molecular approaches to disease gene identification, but overviews and case studies are also presented.

The volume begins with six introductory chapters, which provide overviews of strategies for disease gene identification and functional characterization, and include introductions to microbiome sequencing methods for studying human disease, as well as the emerging role of long noncoding RNAs in human disease. The next section of the text contains chapters presenting methods for identifying potential susceptibility loci, including practical

procedures for genome-wide association analysis, whole genome, whole exome, and single-cell library construction, and methylation profiling.

The volume follows with a section on current applications in human genomics, which provide tools for target validation and functional assessment. These protocols are useful for characterizing disease-associated loci and include methods in quantitative polymerase chain reaction, lentiviral-mediated CRISPR-cas9, RNA interference, and luciferase reporter assay.

We end with four discursive chapters providing examples of disease gene identification and application. The first chapter in this section is related to physiologic interpretation of genome-wide association signals, using type 2 diabetes as a model. The following two chapters present overviews of disease gene discovery in two distinct disorders: hereditary hemochromatosis and small cell carcinoma of the ovary. A discussion of the reemergence of linkage analysis, as an adjunct to association studies, concludes this section.

Completion of this volume would not have been possible without the support and contributions of many individuals. In particular, I thank Dr. John M. Walker, the series editor, who provided expert guidance and oversight at each step of bringing this volume to fruition. I also appreciate the efforts of the authors, all of whom contributed outstanding chapters. It was a pleasure working with this expert team of scientists. It is my hope that this volume leads to the identification and characterization of many more disease-related genes, which may someday pave the way toward more accurate and improved methods for disease diagnosis, as well as novel and effective strategies for disease treatment and prevention.

Phoenix, AZ, USA

Johanna K. DiStefano

Contents

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

PART I INTRODUCTION

1 Identification of Disease Susceptibility Alleles in the Next Generation Sequencing Era	3
<i>Johanna K. DiStefano and Christopher B. Kingsley</i>	
2 Induced Pluripotent Stem Cells in Disease Modeling and Gene Identification	17
<i>Satish Kumar, John Blangero, and Joanne E. Curran</i>	
3 Development of Targeted Therapies Based on Gene Modification	39
<i>Taylor M. Benson, Fatjon Leti, and Johanna K. DiStefano</i>	
4 What Can We Learn About Human Disease from the Nematode <i>C. elegans</i> ?	53
<i>Javier Apfeld and Scott Alper</i>	
5 Microbiome Sequencing Methods for Studying Human Diseases	77
<i>Rebecca M. Davidson and L. Elaine Epperson</i>	
6 The Emerging Role of Long Noncoding RNAs in Human Disease.....	91
<i>Johanna K. DiStefano</i>	

PART II METHODS FOR GENE IDENTIFICATION

7 Identification of Disease-related Genes using a Genome-wide Association Study Approach.....	113
<i>Tobias Wohland and Dorit Schleinitz</i>	
8 Whole Genome Library Construction for Next Generation Sequencing	151
<i>Jonathan J. Keats, Lori Cuyugan, Jonathan Adkins, and Winnie S. Liang</i>	
9 Whole Exome Library Construction for Next Generation Sequencing.....	163
<i>Winnie S. Liang, Kristi Stephenson, Jonathan Adkins, Austin Christofferson, Adrienne Helland, Lori Cuyugan, and Jonathan J. Keats</i>	
10 Optimized Methodology for the Generation of RNA-Sequencing Libraries from Low-Input Starting Material: Enabling Analysis of Specialized Cell Types and Clinical Samples.....	175
<i>Kendra Walton and Brian P. O'Connor</i>	

11 Using Fluidigm C1 to Generate Single-Cell Full-Length cDNA Libraries for mRNA Sequencing 199
Robert Durruthy-Durruthy and Manisha Ray

12 MiSeq: A Next Generation Sequencing Platform for Genomic Analysis 223
Rupesh Kanchi Ravi, Kendra Walton, and Mahdieh Khosroheidari

13 Methods for CpG Methylation Array Profiling Via Bisulfite Conversion 233
Fatjon Leti, Lorida Llaci, Ivana Malenica, and Johanna K. DiStefano

PART III FUNCTIONAL CHARACTERIZATION OF SUSCEPTIBILITY ALLELES AND LOCI

14 miRNA Quantification Method Using Quantitative Polymerase Chain Reaction in Conjunction with C_q Method 257
Fatjon Leti and Johanna K. DiStefano

15 Primary Airway Epithelial Cell Gene Editing Using CRISPR-Cas9 267
Jamie L. Everman, Cydney Rios, and Max A. Seibold

16 RNA Interference to Knock Down Gene Expression 293
Haiyong Han

17 Using Luciferase Reporter Assays to Identify Functional Variants at Disease-Associated Loci 303
Anup K. Nair and Leslie J. Baier

PART IV IDENTIFICATION OF DISEASE GENES

18 Physiologic Interpretation of GWAS Signals for Type 2 Diabetes 323
Richard M. Watanabe

19 Identification of Genes for Hereditary Hemochromatosis 353
Glenn S. Gerhard, Barbara V. Paynton, and Johanna K. DiStefano

20 Identification of Driver Mutations in Rare Cancers: The Role of SMARCA4 in Small Cell Carcinoma of the Ovary, Hypercalcemic Type (SCCOHT) 367
Jessica D. Lang and William P.D. Hendricks

21 The Rise and Fall and Rise of Linkage Analysis as a Technique for Finding and Characterizing Inherited Influences on Disease Expression 381
Ettie M. Lipner and David A. Greenberg

Index 399

Contributors

- JONATHAN ADKINS • *Translational Genomics Research Institute (TGen), Phoenix, AZ, USA*
- SCOTT ALPER • *Department of Biomedical Research, Center for Genes, Environment and Health, National Jewish Health, Denver, CO, USA; Department of Immunology and Microbiology, University of Colorado School of Medicine, Aurora, CO, USA*
- JAVIER APFELD • *Department of Biology, Northeastern University, Boston, MA, USA*
- LESLIE J. BAIER • *Diabetes Molecular Genetics Section, Phoenix Epidemiology and Clinical Research Branch, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Phoenix, AZ, USA*
- TAYLOR M. BENSON • *Department of Biomedical Research, Center for Genes, Environment, and Health, National Jewish Health, Denver, CO, USA*
- JOHN BLANGERO • *South Texas Diabetes and Obesity Institute, School of Medicine, University of Texas Rio Grande Valley, Edinburg, TX, USA*
- AUSTIN CHRISTOFFERSON • *Translational Genomics Research Institute (TGen), Phoenix, AZ, USA*
- JOANNE E. CURRAN • *South Texas Diabetes and Obesity Institute, School of Medicine, University of Texas Rio Grande Valley, Edinburg, TX, USA*
- LORI CUYUGAN • *Translational Genomics Research Institute (TGen), Phoenix, AZ, USA*
- REBECCA M. DAVIDSON • *Department of Biomedical Research, Center for Genes, Environment and Health, National Jewish Health, Denver, CO, USA*
- JOHANNA K. DiSTEFANO • *Translational Genomics Research Institute, Phoenix, AZ, USA*
- ROBERT DURRUTHY-DURRUTHY • *Fluidigm Corporation, South San Francisco, CA, USA*
- L. ELAINE EPPERSON • *Department of Biomedical Research, Center for Genes, Environment and Health, National Jewish Health, Denver, CO, USA*
- JAMIE L. EVERMAN • *Center for Genes, Environment, and Health, National Jewish Health, Denver, CO, USA*
- GLENN S. GERHARD • *Department of Medical Genetics and Molecular Biochemistry, 960 Medical Education and Research Building (MERB), Lewis Katz School of Medicine at Temple University, Philadelphia, PA, USA*
- DAVID A. GREENBERG • *Battelle Center for Mathematical Medicine, Nationwide Children's Hospital, Columbus, OH, USA; Department of Pediatrics, Wexner Medical Center, Ohio State University, Columbus, OH, USA*
- HAIYONG HAN • *Translational Genomics Research Institute, Phoenix, AZ, USA*
- ADRIENNE HELLAND • *Translational Genomics Research Institute (TGen), Phoenix, AZ, USA*
- WILLIAM P.D. HENDRICKS • *Translational Genomics Research Institute, Phoenix, AZ, USA*
- JONATHAN J. KEATS • *Translational Genomics Research Institute (TGen), Phoenix, AZ, USA*
- MAHDIEH KHOSROHEIDARI • *Illumina, Inc., San Diego, CA, USA*
- CHRISTOPHER B. KINGSLEY • *Roche Sequencing Solutions, Inc., San Jose, CA, USA*
- SATISH KUMAR • *South Texas Diabetes and Obesity Institute, School of Medicine, University of Texas Rio Grande Valley, Edinburg, TX, USA*
- JESSICA D. LANG • *Translational Genomics Research Institute, Phoenix, AZ, USA*
- FATJON LETI • *Department of Biomedical Research, Center for Genes, Environment, and Health, National Jewish Health, Denver, CO, USA*

- WINNIE S. LIANG • *Translational Genomics Research Institute (TGen), Phoenix, AZ, USA*
- ETTIE M. LIPNER • *Center for Genes, Environment, and Health, National Jewish Health, Denver, CO, USA; Department of Pharmacology, School of Medicine, University of Colorado Denver, Aurora, CO, USA*
- LORIDA LLACI • *Neurogenomics Division, Translational Genomics Research Institute, Phoenix, AZ, USA*
- IVANA MALENICA • *Division of Biostatistics, University of California, Berkeley, Berkeley, CA, USA*
- ANUP K. NAIR • *Diabetes Molecular Genetics Section, Phoenix Epidemiology and Clinical Research Branch, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Phoenix, AZ, USA*
- BRIAN P. O'CONNOR • *Genomics Facility, Center for Genes, Environment and Health, National Jewish Health, Denver, CO, USA*
- BARBARA V. PAYNTON • *Department of Medical Genetics and Molecular Biology, The Lewis Katz School of Medicine at Temple University, Philadelphia, PA, USA*
- RUPESH KANCHI RAVI • *Pfizer, Inc., San Diego, CA, USA*
- MANISHA RAY • *Fluidigm Corporation, South San Francisco, CA, USA*
- CYDNEY RIOS • *Center for Genes, Environment, and Health, National Jewish Health, Denver, CO, USA*
- DORIT SCHLEINITZ • *IFB Adiposity Diseases, Leipzig University Medical Center, University of Leipzig - Medical Faculty, Leipzig, Germany; Clinic and Policlinic for Endocrinology and Nephrology, Leipzig University Medical Center, Leipzig, Germany*
- MAX A. SEIBOLD • *Center for Genes, Environment, and Health, National Jewish Health, Denver, CO, USA; Department of Pediatrics, National Jewish Health, Denver, CO, USA*
- KRISTI STEPHENSON • *Translational Genomics Research Institute (TGen), Phoenix, AZ, USA*
- KENDRA WALTON • *Genomics Facility, Center for Genes, Environment, and Health, National Jewish Health, Denver, CO, USA*
- RICHARD M. WATANABE • *Departments of Preventive Medicine and Physiology & Biophysics, Keck School of Medicine of USC, Los Angeles, CA, USA*
- TOBIAS WOHLAND • *IFB Adiposity Diseases, Leipzig University Medical Center, University of Leipzig - Medical Faculty, Leipzig, Germany*

Part I

Introduction

Chapter 1

Identification of Disease Susceptibility Alleles in the Next Generation Sequencing Era

Johanna K. DiStefano and Christopher B. Kingsley

Abstract

The development of next generation sequencing (NGS) technologies has transformed the study of human genetic variation. In less than a decade, NGS has facilitated the discovery of causal mutations in both rare, monogenic diseases and common, heterogeneous disorders, leading to unprecedented improvements in disease diagnosis and treatment strategies. Given the rapid evolution of NGS platforms, it is now possible to analyze whole genomes and exomes quickly and affordably. Further, emerging NGS applications, such as single-cell sequencing, have the power to address specific issues like somatic variation, which is yielding new insights into the role of somatic mutations in cancer and late-onset diseases. Despite limitations associated with current iterations of NGS technologies, the impact of this approach on identifying disease-causing variants has been significant. This chapter provides an overview of several NGS platforms and applications and discusses how these technologies can be used in concert with experimental and computational strategies to identify variants with a causative effect on disease development and progression.

Key words Genetics, Human disease, Next generation sequencing, Causal variant, Whole genome sequencing, Exome sequencing, Single cell sequencing, Somatic mosaicism

1 Introduction

DNA sequence variation is a common feature of all organisms. In humans, approximately 0.1–0.4% of nucleotides differ between any given pair of unrelated genomes [1, 2]. The vast majority of sequence variation is comprised of single nucleotide variants (SNVs), which occur every 100–300 bases [3], and are mostly located within noncoding sequence [4]. A large number of inherited human diseases are caused by sequence variation in single genes [5–7], and many complex diseases, including cancer, diabetes, and heart disease, are mediated, at least in part, by genetic factors [8–11]. The majority of rare diseases, such as those affecting only a small percentage of the population, result from hereditary or de novo genetic mutations [12]. In recent years, significant effort has

been made to identify and characterize causal variants underlying a vast spectrum of human diseases.

Technological advances in high-throughput genotyping methods over the past two decades revolutionized the field of human genetics. In particular, cost-effective, microarray-based genotyping of vast numbers of SNVs enabled population-based, genome-wide association studies for common human diseases such as diabetes, neurological disorders, and cancer [13–17]. Genome-wide association approaches have identified statistically significant evidence supporting relationships between complex human diseases and hundreds of common genetic variants in the human population. However, finding disease-associated alleles is only the first step on the path to identifying those variants that directly contribute to disease risk. A major challenge inherent in these studies is moving from identification of a genetic variant via association studies to determination of actual causal variants through functional genomics experimentation.

In the past, identification of causal variants involved querying public databases for the presence of characterized functional elements in the vicinity of associated alleles. These functional elements would then be prioritized for targeted resequencing of relatively small genomic regions in affected individuals, with the goal of identifying novel variants that directly impact disease development. The success of such an approach depended on the means and accuracy with which functional elements were identified, and traditional low-throughput sequencing technologies placed severe limitations on the number of individuals and the span of genomic regions that could be resequenced in an economically feasible way.

Many “next generation” technologies have emerged over the past several years, promising inexpensive and efficient sequencing of large amounts of genomic DNA as an alternative to microarray genotyping studies [18]. Next generation sequencing (NGS) refers to a field of technologies geared toward massively parallel sequence analysis of nucleic acids. Compared with traditional Sanger sequencing, NGS provides faster and cheaper analysis of samples and yields a much greater amount of sequence information for each individual sample [19, 20]. Advances in NGS technology have removed the obstacles associated with low throughput, labor-intensive, and expensive traditional sequencing, and is now enabling researchers to identify many more novel sequence variants than previously possible.

In this chapter, we briefly address popular commercial NGS platforms and provide an overview of several NGS methods for the identification of disease genes, including whole genome sequencing, exome sequencing, RNA-sequencing, and single-cell genome sequencing, all of which complement the methods-based descriptions found in Chapters 8–11. We also discuss the manner in which

NGS technologies can be used in concert with experimental and computational strategies to identify alleles that contribute to the development of human disease.

2 Methods

2.1 Next Generation Sequencing (NGS)

Since the Watson–Crick publication of the three-dimensional structure of DNA in 1953 [21], a number of different techniques and technologies have been applied to nucleic acid sequence elucidation (Fig. 1), which has led to a revolution in the biological sciences. The whole genome sequences that began to appear in the 1990s were made possible only by the development of commercial instruments based upon the capillary electrophoresis-based Sanger sequencing method [22, 23]. These advances paved the way for the birth of entire research disciplines devoted to the analysis of the large amount of nucleotide sequence. An excellent review by Heather and Chain [24] describes the history of DNA sequencing, from initial efforts in RNA sequencing to emerging “third-generation sequencing” approaches, giving a thorough historical perspective of the changes in technologies that have occurred over the years.

2.1.1 Currently Popular NGS Platforms

NGS has emerged as a new set of technologies that have not only reduced the cost of sequencing entire mammalian genomes from many millions of dollars to approximately one thousand dollars, but also decreased the amounts of time and effort needed to complete the task. NGS technologies have been, and continue to be, developed by a number of commercial entities, and at the time of this writing, include those that utilize (1) fluorescent imaging such as the Illumina NGS system and the Pacific Biosciences SMRT NGS platform; (2) pH (i.e., Ion Torrent); and (3) nanopore technology (i.e., Oxford Nanopore and Genia [Roche]). The Illumina platform uses sequencing by synthesis (SBS) technology in which sequential

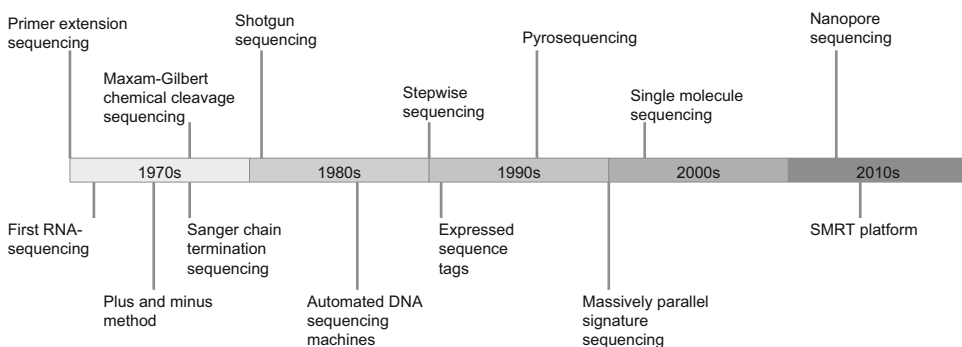


Fig. 1 Timeline of technological developments in nucleic acid sequencing

fluorophore-labeled nucleotide base addition combined with fluorescence imaging determines the identity of the incorporated nucleotide. This technology also allows DNA fragments to be sequenced from both ends (i.e., pair-end sequencing), which enhances detection of rearrangements (i.e., insertions, deletions, and inversions), repetitive sequence elements, gene fusions, and novel transcripts [25–27]. The SMRT NGS platform utilizes zero-mode waveguides (ZMWs), which strategically permit illumination only of the bottom of the well, where a DNA polymerase/template is immobilized. Nucleotides, each labeled with a uniquely colored fluorophore, are exposed to the template, and when one is held in position, it emits a light pulse. During incorporation the phosphate chain is cleaved, releasing the fluorophore. Because SMRT sequencing does not pause after nucleotide incorporation for chemical cleavage and fluorescence imaging, this technology is much faster than the SBS technology of Illumina. The Ion Torrent takes advantage of the naturally occurring proton release that occurs when a nucleotide is incorporated into a strand of DNA. In each sequencing step, only a single nucleotide type is introduced into the reaction flow cell, so only the DNA fragment with the corresponding nucleotide as the next base will show a change in pH signal. The nanopore approach differs from the preceding platforms by pulling target DNA molecules through nanopores embedded in synthetic polymer membranes. The Oxford Nanopore pulls target DNA molecules through a synthetic polymer membrane, which elicits an electric current affected by the identity of the nucleotides in the pore complex at that moment. The Genia system incorporates a DNA polymerase tethered to the pore complex and specifically modified dNTPs using nano-tags specific for each nucleotide. As a primer is extended off the target DNA template, the tag is cleaved and flows through the nanopore, inducing a change in electric current. An in-depth review of platforms, as well as emerging NGS technologies, can be found elsewhere [18].

2.1.2 Whole Genome Sequencing

Despite the success of genome-wide association studies (GWAS) in enhancing understanding of disease mechanisms, the variants identified by this approach represent only a fraction of the overall genetic contribution to common disease risk. While many disease-associated variants have been identified through GWAS, they have mostly been common variants with moderate to high (i.e., >0.1) allele frequencies and moderate to low (i.e., 1.1–1.3) odds ratios. The accumulated results of many GWAS have called into question the validity of the “common disease-common variant” hypothesis for complex diseases; the proposal that common polymorphisms contribute to a significant proportion of the susceptibility to common diseases [28–30].

In the face of these findings, some researchers have argued that common variants with low odds ratios are unlikely to be responsible for the observed familial clustering of many common diseases, such as heart disease and diabetes, because familial clustering generally requires that individuals who share the risk allele have a high probability of displaying the phenotype. This has led some to hypothesize that the majority of the inherited risk is caused by a heterogeneous collection of rare variants that exist in the population [31, 32]. While identification of these rare variants has become a major focus in human genetics, study samples of tens of thousands of carefully phenotyped and appropriately stratified individuals are required to adequately power a GWAS to identify them [33]. Even under these conditions, this approach still fails to account for the role of environmental factors in disease susceptibility, which in complex disorders like diabetes and heart disease may be more impactful than genetic factors [31].

In defense of the common disease-common variant hypothesis, however, one recent large scale sequencing study involving T2D patients found that variants associated with that disease were overwhelmingly common, and most fell in regions that had been previously identified by GWAS [34]. Additional studies will be needed to validate this result and extend it to other diseases, but the question of whether common or rare variants underlie the majority of risk for common diseases continues to remain an open one.

Because of the limited success of GWAS, most researchers have now turned to whole-genome sequencing (WGS) to identify rare causal variants. WGS provides an unbiased analysis of the entire genome, including potential, not-yet-annotated genes, noncoding RNAs, and regulatory regions. In contrast, whole exome sequencing (WES), discussed in the following section, focuses on analysis of the protein-coding genome. Costs associated with WGS are higher than those of WES; however, costs for WGS are decreasing more rapidly than WES, and will likely approximate WES costs in the near future. A recent comparison of the two methods in six individuals found the distribution of sequencing quality parameters, including the number of aligned reads covering a single position (coverage depth), genotype quality, and the ratio of reads for the minor allele (minor-read ratio) for SNVs and insertion-deletions (indels) to be more uniform for WGS [35]. Differences were attributed to effects resulting from the hybridization/capture and PCR-amplification steps required for the preparation of WES sequencing libraries. Of note, approximately 650 high-quality coding variants were identified by WGS, but missed by WES, suggesting that the former method may be better for detecting exomic mutations.

WGS has been used to identify a number of mutations with direct causal effects on diseases such as amyotrophic lateral sclerosis [36], retinitis pigmentosa [37], dystonia [38], and autism spectrum

disorder [39]. Large-scale WGS was recently performed in ~2600 Icelanders [40]. Using a combination of filtered sequence data (excess of homozygosity, rare, protein-coding), imputation of these variants into >100,000 Icelanders, and GWAS, the investigators identified a recessive frameshift mutation in myosin light chain 4 (*MYL4*) that causes early-onset atrial fibrillation. This study provides a strong paradigm by which WGS data can be used in conjunction with imputation and association analysis to identify disease-causing variants. The results also underscore how such a design may yield a better understanding of the role of genomic sequence variation in human diversity.

2.1.3 Whole Exome Sequencing

WES utilizes capture of all coding regions prior to selective analysis of the exomic content of the genome [41]. One lesson learned from GWAS is that the genetic landscape for common diseases is often quite complex, with many variants of small effect interacting with each other and the environment to determine the risk for an individual. With this in mind, many researchers turned to a simpler case of identifying and sequencing individuals who suffer from rare, extreme phenotypes that resemble a common disease [42]. Rare or sporadic cases such as these can be the result of one or more low-frequency, disease-causing variants segregating in a Mendelian family, but are also often a result of de novo mutations in an individual, especially dominant or haploinsufficient mutations.

Rare diseases are generally the result of a small number (often one) of mutations with large effects. Such variants often stand out; they tend to occur in or near coding regions, with obvious effects such as missense, nonsense, or splice site mutations. Indeed, the majority of all disease-causing mutations in monogenic disorders are located in the protein-coding genome [43]. Not surprisingly, sequence analysis of monogenic diseases is more straightforward compared to diseases of complex etiology. The first successful application of WES was used to identify the genetic basis of Miller syndrome, which is characterized by severe micrognathia, cleft lip/palate, hypoplasia or aplasia of the posterior elements of the limbs, coloboma of the eyelids, and supernumerary nipples [44]. In that study, the investigators captured and sequenced coding regions of two affected siblings and two affected unrelated individuals, and after variant filtering and use of algorithms to assess whether a mutation was damaging, identified dihydroorotate dehydrogenase (*DHODH*) as the disease gene. This study was the first to show that WES of a small number of unrelated, affected individuals is an efficient strategy to identify genes underlying rare monogenic diseases. Since this initial study, nearly 1000 novel monogenic disease genes have been identified using a similar design [45].

In our own unpublished studies, we applied a WES approach to identify novel genetic mutations underpinning short QT syndrome

(SQTS) in a Native American pedigree with a high incidence of ventricular arrhythmias and sudden cardiac death. SQTS is a cardiac electrical disorder characterized by an abnormally short QT interval and associated with an increased risk of atrial and ventricular tachyarrhythmias and sudden cardiac death [46]. Six distinct subtypes of SQTS have been linked to six genes, although the majority of genetic factors underpinning the disease remain unknown. We investigated a family with a high incidence of ventricular arrhythmias and sudden cardiac death, in which the disease follows an autosomal dominant pattern of inheritance. Genetic analysis excluded known SQTS variants in this family. To identify the mutation underlying the disease in this pedigree, we performed WES for two affected siblings using the SureSelect platform and the HiSeq2000 sequencing system. We first filtered variants by mutation type and presence in both siblings and found 1881 entries. To prioritize variants we used an integrated knowledge mining approach to identify genes and biological concepts associated with SQTS. Of these, we selected several hundred variants for genotyping using the MiSeq platform in four SQTS cases, four unaffected controls, and two individuals of unknown disease status. Out of the 788 loci analyzed, we found only five heterozygous variants that exhibited complete sharing in cases and were absent in unaffected controls. Three of the variants were present as missense mutations with $MAF \leq 0.006$ in the 1000 Genomes database. The fourth variant was intronic and also present in 1000 Genomes ($MAF = 0.017$). The remaining variant was a novel exonic C/T substitution in the cystic fibrosis transmembrane conductance regulator gene (*CFTR*, chr7: 117267757 [C/T]), which was predicted to be a stop-gain mutation by wANNOVAR (<http://wannovar.usc.edu/>) and “disease causing” by MutationTaster (<http://mutationtaster.org>). Although all five mutations may potentially contribute to disease risk, the *CFTR* mutation warrants further attention due to its absence from the public databases and its potential deleterious effects on ion conductance in cardiac tissue.

2.1.4 Single-Cell Sequencing

Recent innovations in cell isolation techniques, NGS, and bioinformatics analytical methods now enable WGS of a single-cell genome. As its name implies, single-cell sequencing analyzes the genomes of individual cells. Single cells of interest are isolated from suspension using any number of different methods [47] and the single copy of the genome is amplified using techniques such as multiplex displacement amplification [48]. WGS of the amplified product has advanced the ability to observe mosaicism at the cellular level. For example, a trio of recent studies applying single-cell sequencing to neurons revealed that mosaicism is a widespread phenomenon in the brain. The first study used WGS of single neurons from a healthy individual to identify spontaneous somatic mutations as

clonal marks to track cell lineages in human brain [49]. Lovato et al. [50] used single-cell whole-genome sequencing of 36 neurons from the cerebral cortex of three normal individuals to identify thousands of mosaic SNVs. The results indicated that neuronal mutations are enriched at sites of active transcription, in contrast to germline and cancer SNVs, which are typically acquired during DNA replication. The authors also performed single-cell WGS in >160 neurons from three normal and two pathological human brains and found $\geq 95\%$ of neurons in normal brain tissue to be euploid. In a patient with hemimegalencephaly, a rare neurological condition in which one half of the brain is much larger than the other [51], due to somatic copy number variation (CNV) on chromosome 1q, the investigators found unexpected tetrasomy 1q in approximately 20% of neurons, suggesting that CNVs in a minority of cells can cause widespread brain dysfunction. Single-cell analyses thus revealed the presence of large private and clonal somatic CNVs in both normal and diseased human brains.

The human genome acquires mutations spontaneously with cell division [52], which contribute to the development of human diseases [53, 54], including cancer [55]. As experimental and analytical advances continue to reduce the technical noise from single-cell WGA, which will improve resolution of true variants from experimental artifacts, appreciation for the role of low-level mosaic genetic variants in the etiology of human disease is expected to grow.

2.1.5 *Limitations of NGS*

NGS continues to be an evolving field and is advancing so rapidly that almost any review article on the subject will be outdated by the time of publication. However, while many advances have been made, particularly in the last 5 years, many challenges associated with the analysis of NGS have also emerged. From a technical standpoint, longer read lengths and lower error rates will improve the accuracy of alignment of sequences to the reference genome, with subsequent increases in the sensitivity and specificity of detecting genuine sequence variants. Analytically, improved algorithms for variant detection, especially structural variants such as insertion–deletion mutations, will more completely identify sequence variants in genomic regions of interest. Time and cost are also issues that must be considered with each platform. We can expect that further refinements of the sequencing technology leading to lower error rates, longer read lengths, and faster turnover, combined with improved computational methods, will be effective in overcoming these difficulties to the point where a small research group or academic laboratory will be able to sequence large genomic regions, or even entire genomes, in a matter of hours for only a few hundred dollars.

From an analytical standpoint, the sheer volume of variants detected with NGS represents a significant challenge. For example, 20 million SNVs and 1.5 million insertions-deletions were identified in a recent WGS study [40]. WES typically identifies ~20,000 SNVs per genome [45], although following filtering for potentially deleterious variants, a single exome is estimated to have 100–200 potential disease-causing mutations [56]. Current figures estimate that the average human carries ~100 mutations that cause loss-of-function within protein-coding genes [57], although these variants do not cause disease-related phenotypes [58]. Therefore, at this time, the analysis of enormous amounts of sequence data and functional validation of potential disease-causing mutations, as opposed to mutation detection, represent the major bottlenecks for translating sequence information into clinical practice.

2.2 Post-sequencing Prioritization of Potential Variants

Once a list of variants has been assembled from a WGS/WES study, the next step is to prioritize them for experimental assays that assess functional consequences. Prioritization of variants is typically performed through a sequential set of filters until the number of variants to be tested is reduced to a manageable size, given the available experimental resources and appropriate experimental approach. These filters often involve statistical analysis of genetic association in primary and validation study samples, followed by prediction of the functional consequences of individual sequence variants, which makes use of genome annotation information.

The first selection criterion usually applied to presumptive causal sequence variants is the strength of the association between individual markers and the trait of interest in the original study sample. Those markers showing the strongest evidence for statistical significance (i.e., lowest p -value or highest odds ratio) in the discovery study sample are selected for validation in a second, preferably independent, population to reduce the number of falsely positive results. While validation samples are usually drawn from populations of the same ethnicity as the discovery sample, different ethnic populations can be used when many adjacent markers are significantly associated with the outcome due to high levels of correlation between markers (i.e., linkage disequilibrium). In these cases, validation in a sample of a different ethnic background can sometimes distinguish causal variants from indirectly associated variants because of the different allele frequencies and patterns of linkage disequilibrium among ethnic populations. For example, validation in samples of African origin can refine genetic associations due to shorter blocks of linkage disequilibrium relative to populations of other ethnic backgrounds, in whom linkage disequilibrium typically spans greater intervals. This strategy can be complicated, however, in that population differences in allele frequencies may also lead to decreased power to detect genuine genetic associations in the validation population [59].

Analysis of the genomic context of the associated sequence variants is typically the second step in marker prioritization. Variants with clear functional consequences on coding sequence (e.g., nonsense, missense, and splice site mutations) are obvious candidates for further investigation. In addition, analytical approaches have been developed to predict the functional consequences of nonsynonymous mutations based upon the analysis of multiple sequence alignments and/or protein three-dimensional structure [60]. For those cases in which the associated variants occur far from any known genes, as has been observed for most regions identified by GWAS, genome annotation information can be very useful. For this class of variants, colocalization with functional genomic elements such as transcription factor binding sites, noncoding RNAs, and regions of strong phylogenetic conservation can be taken into consideration when prioritizing potential causal variants for downstream molecular characterization.

By applying genetic and genomic filters such as those described above, the number of associated sequence variants can be pared down to a reasonable size for *in vitro* and *in vivo* functional studies to assess the effect of the variant on some qualitative or quantitative outcome. For those variants that affect coding sequences of genes whose protein products possess measurable activity, this can be a straightforward process of expressing a version of the protein containing the variant and conducting the appropriate assay. For those variants that occur in gene-proximal regulatory elements, transcriptional effects can be measured using reporter constructs containing the variant compared with the normal sequence in transfection experiments [61]. For those variants that occur far from known genes, transfection experiments may also be used to measure transcriptional effects, although transgenic and knockout technology in mice has also been used, and may be more appropriate for long-range acting regulatory sequences [62].

3 Conclusions

Massively parallel DNA sequencing capability, available with current NGS platforms, has revolutionized the field of human genetics. Up until recently, WGS was not accessible to most individual research groups, but declining costs and greater availability of NGS platforms through core services and commercial entities now permits widespread use of this technology. NGS applications have thus far made the greatest impact in the area of monogenic disorders, through the identification of novel determinants of disease and elucidating new pathways important for normal biology. These advances not only improve management of rare disorders but also enhance understanding of critical pathogenic mechanisms underlying common, complex diseases.

References

1. Jorde LB, Wooding SP (2004) Genetic variation, classification and 'race'. *Nat Genet* 36: S28–S33
2. Tishkoff SA, Kidd KK (2004) Implications of biogeography of human populations for 'race' and medicine. *Nat Genet* 36:S21–S27
3. Ke X, Taylor MS, Cardon LR (2008) Singleton SNPs in the human genome and implications for genome-wide association studies. *Eur J Hum Genet* 16:506–515
4. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D, S. N. P. M. W. G. International (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–933
5. Arredondo-Vega FX, Santisteban I, Daniels S, Toutain S, Hershfield MS (1998) Adenosine deaminase deficiency: genotype-phenotype correlations based on expressed activity of 29 mutant alleles. *Am J Hum Genet* 63:1049–1059
6. Bobadilla JL, Macek M Jr, Fine JP, Farrell PM (2002) Cystic fibrosis: a worldwide analysis of CFTR mutations--correlation with incidence data and application to screening. *Hum Mutat* 19:575–606
7. Walker FO (2007) Huntington's disease. *Lancet* 369:218–228
8. Encinas G, Maistro S, Pasini FS, Katayama ML, Brentani MM, Bock GH, Folgueira MA (2015) Somatic mutations in breast and serous ovarian cancer young patients: a systematic review and meta-analysis. *Rev Assoc Med Bras* 61:474–483
9. Kaul N, Ali S (2016) Genes, genetics, and environment in type 2 diabetes: implication in personalized medicine. *DNA Cell Biol* 35:1–12
10. Orho-Melander M (2015) Genetics of coronary heart disease: towards causal mechanisms, novel drug targets and more personalized prevention. *J Intern Med* 278:433–446
11. Srivastava I, Thukral N, Hasiya Y (2015) Genetics of human age related disorders. *Adv Gerontol* 28:228–247
12. Puiu M, Dan D (2010) Rare diseases, from European resolutions and recommendations to actual measures and strategies. *Maedica (Buchar)* 5:128–131
13. Chen H, Yu H, Wang J, Zhang Z, Gao Z, Chen Z, Lu Y, Liu W, Jiang D, Zheng SL, Wei GH, Issacs WB, Feng J, Xu J (2015) Systematic enrichment analysis of potentially functional regions for 103 prostate cancer risk-associated loci. *Prostate* 75:1264–1276
14. Karaderi T, Drong AW, Lindgren CM (2015) Insights into the genetic susceptibility to type 2 diabetes from genome-wide association studies of obesity-related traits. *Curr Diab Rep* 15:83
15. Mocellin S, Verdi D, Pooley KA, Nitti D (2015) Genetic variation and gastric cancer risk: a field synopsis and meta-analysis. *Gut* 64:1209–1219
16. Neale BM, Sklar P (2015) Genetic analysis of schizophrenia and bipolar disorder reveals polygenicity but also suggests new directions for molecular interrogation. *Curr Opin Neurobiol* 30:131–138
17. Peters U, Bien S, Zubair N (2015) Genetic architecture of colorectal cancer. *Gut* 64:1623–1636
18. Khodakov D, Wang C, Zhang DY (2016) Diagnostics based on nucleic acid sequence variant profiling: PCR, hybridization, and NGS approaches. *Adv Drug Deliv Rev* 105:3
19. Buermans HP, den Dunnen JT (2014) Next generation sequencing technology: advances and applications. *Biochim Biophys Acta* 1842:1932–1941
20. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. *Trends Genet* 30:418–426
21. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171:737–738
22. Hunkapiller T, Kaiser RJ, Koop BF, Hood L (1991) Large-scale and automated DNA sequence determination. *Science* 254:59–67
23. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (1977) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265:687–695
24. Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. *Genomics* 107:1–8
25. Bashir A, Volik S, Collins C, Bafna V, Raphael BJ (2008) Evaluation of paired-end sequencing

- strategies for detection of genome rearrangements in cancer. *PLoS Comput Biol* 4: e1000051
26. Cao MD, Tasker E, Willadsen K, Imelfort M, Vishwanathan S, Sureshkumar S, Balasubramanian S, Boden M (2014) Inferring short tandem repeat variation from paired-end short reads. *Nucleic Acids Res* 42:e16
 27. Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30:614–620
 28. Chakravarti A (1999) Population genetics--making sense out of sequence. *Nat Genet* 21:56–60
 29. Lander ES (1996) The new genomics: global views of biology. *Science* 274:536–539
 30. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273:1516–1517
 31. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. *Nature* 461:747–753
 32. Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* 19:212–219
 33. Lee S, Abecasis GR, Boehnke M, Lin X (2014) Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* 95:5–23
 34. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, Ma C, Fontanillas P, Moutsianas L, McCarthy DJ, Rivas MA, Perry JR, Sim X, Blackwell TW, Robertson NR, Rayner NW, Cingolani P, Locke AE, Tajos JF, Highland HM, Dupuis J, Chines PS, Lindgren CM, Hartl C, Jackson AU, Chen H, Huyghe JR, van de Bunt M, Pearson RD, Kumar A, Muller-Nurasyid M, Grarup N, Stringham HM, Gamazon ER, Lee J, Chen Y, Scott RA, Below JE, Chen P, Huang J, Go MJ, Stitzel ML, Pasko D, Parker SC, Varga TV, Green T, Beer NL, Day-Williams AG, Ferreira T, Fingerlin T, Horikoshi M, Hu C, Huh I, Ikram MK, Kim BJ, Kim Y, Kim YJ, Kwon MS, Lee J, Lee S, Lin KH, Maxwell TJ, Nagai Y, Wang X, Welch RP, Yoon J, Zhang W, Barzilai N, Voight BF, Han BG, Jenkinson CP, Kuulusmaa T, Kuusisto J, Manning A, Ng MC, Palmer ND, Balkau B, Stancakova A, Abboud HE, Boeing H, Giedraitis V, Prabhakaran D, Gottesman O, Scott J, Carey J, Kwan P, Grant G, Smith JD, Neale BM, Purcell S, Butterworth AS, Howson JM, Lee HM, Lu Y, Kwak SH, Zhao W, Danesh J, Lam VK, Park KS, Saleheen D, So WY, Tam CH, Afzal U, Aguilar D, Arya R, Aung T, Chan E, Navarro C, Cheng CY, Palli D, Correa A, Curran JE, Rybin D, Farook VS, Fowler SP, Freedman BI, Griswold M, Hale DE, Hicks PJ, Khor CC, Kumar S, Lehne B, Thuillier D, Lim WY, Liu J, van der Schouw YT, Loh M, Musani SK, Puppala S, Scott WR, Yengo L, Tan ST, Taylor HA Jr, Thameem F, Wilson G, Wong TY, Njolstad PR, Levy JC, Mangino M, Bonnycastle LL, Schwarzmayr T, Fadista J, Surdulescu GL, Herder C, Groves CJ, Wieland T, Bork-Jensen J, Brandslund I, Christensen C, Koistinen HA, Doney AS, Kinnunen L, Esko T, Farmer AJ, Hakaste L, Hodgkiss D, Kravic J, Lysenko V, Hollensted M, Justesen ME, Jorgensen T, Ladvall C, Jurgens JM, Karajamaki A, Kriebel J, Rathmann W, Lannfelt L, Lauritzen T, Narisu N, Linneberg A, Melander O, Milani L, Neville M, Orholm Melander M, Qi L, Qi Q, Roden M, Rolandsson O, Swift A, Rosengren AH, Stirrups K, Wood AR, Mihailov E, Blancher C, Carneiro MO, Maguire J, Poplin R, Shakir K, Fennell T, DePristo M, Hrade de Angelis M, Deloukas P, Gjesing AP, Jun G, Nilsson P, Murphy J, Onofrio R, Thorand B, Hansen T, Meisinger C, Hu FB, Isomaa B, Karpe F, Liang L, Peters A, Huth C, O'Rahilly SP, Palmer CN, Pedersen O, Rauramaa R, Tuomilehto J, Salomaa V, Watanabe RM, Syvanen AC, Bergman RN, Bharadwaj D, Bottinger EP, Cho YS, Chandak GR, Chan JC, Chia KS, Daly MJ, Ebrahim SB, Langenberg C, Elliott P, Jablonski KA, Lehman DM, Jia W, Ma RC, Pollin TI, Sandhu M, Tandon N, Froguel P, Barroso I, Teo YY, Zeggini E, Loos RJ, Small KS, Ried JS, DeFronzo RA, Grallert H, Glaser B, Metspalu A, Wareham NJ, Walker M, Banks E, Gieger C, Ingelsson E, Im HK, Illig T, Franks PW, Buck G, Trakalo J, Buck D, Prokopenko I, Magi R, Lind L, Farjoun Y, Owen KR, Gloy AL, Strauch K, Tuomi T, Koener JS, Lee JY, Park T, Donnelly P, Morris AD, Hattersley AT, Bowden DW, Collins FS, Atzmon G, Chambers JC, Spector TD, Laakso M, Strom TM, Bell GI, Blangero J, Duggirala R, Tai ES, McVean G, Hanis CL, Wilson JG, Seielstad M, Frayling TM, Meigs JB, Cox NJ, Sladek R, Lander ES, Gabriel S, Burt NP, Mohlke KL, Meitinger T, Groop L, Abecasis G, Florez JC, Scott LJ, Morris AP, Kang HM, Boehnke M,

- Altshuler D, McCarthy MI (2016) The genetic architecture of type 2 diabetes. *Nature* 536:41
35. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, Shang L, Boisson B, Casanova JL, Abel L (2015) Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A* 112:5473–5478
 36. Herdewyn S, Zhao H, Moisse M, Race V, Matthijs G, Reumers J, Kusters B, Schelhaas HJ, van den Berg LH, Goris A, Robberecht W, Lambrechts D, Van Damme P (2012) Whole-genome sequencing reveals a coding non-pathogenic variant tagging a non-coding pathogenic hexanucleotide repeat expansion in C9orf72 as cause of amyotrophic lateral sclerosis. *Hum Mol Genet* 21:2412–2419
 37. Nishiguchi KM, Tearle RG, Liu YP, Oh EC, Miyake N, Benaglio P, Harper S, Koskiniemi-Kuendig H, Venturini G, Sharon D, Koenekoop RK, Nakamura M, Kondo M, Ueno S, Yasuma TR, Beckmann JS, Ikegawa S, Matsumoto N, Terasaki H, Berson EL, Katsanis N, Rivolta C (2013) Whole genome sequencing in patients with retinitis pigmentosa reveals pathogenic DNA structural changes and NEK2 as a new disease gene. *Proc Natl Acad Sci U S A* 110:16139–16144
 38. Lohmann K, Wilcox RA, Winkler S, Ramirez A, Rakovic A, Park JS, Arns B, Lohnau T, Groen J, Kastan M, Bruggemann N, Hagenah J, Schmidt A, Kaiser FJ, Kumar KR, Zschiedrich K, Alvarez-Fischer D, Altenmuller E, Ferbert A, Lang AE, Munchau A, Kostic V, Simonyan K, Agzarian M, Ozelius LJ, Langeveld AP, Sue CM, Tijssen MA, Klein C (2013) Whispering dysphonia (DYT4 dystonia) is caused by a mutation in the TUBB4 gene. *Ann Neurol* 73:537–545
 39. Jiang YH, Yuen RK, Jin X, Wang M, Chen N, Wu X, Ju J, Mei J, Shi Y, He M, Wang G, Liang J, Wang Z, Cao D, Carter MT, Chrysler C, Drmic IE, Howe JL, Lau L, Marshall CR, Merico D, Nalpathamkalam T, Thiruvahindrapuram B, Thompson A, Uddin M, Walker S, Luo J, Anagnostou E, Zwaigenbaum L, Ring RH, Wang J, Lajonchere C, Wang J, Shih A, Szatmari P, Yang H, Dawson G, Li Y, Scherer SW (2013) Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am J Hum Genet* 93:249–263
 40. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, Sigurdsson GT, Stacey SN, Frigge ML, Holm H, Saemundsdottir J, Helgadóttir HT, Johannsdóttir H, Sigfusson G, Thorgeirsson G, Sverrisson JT, Gretarsdóttir S, Walters GB, Rafnar T, Thjodleifsson B, Bjornsson ES, Olafsson S, Thorarinsdóttir H, Steingrimsdóttir T, Gudmundsdóttir TS, Theodors A, Jonasson JG, Sigurdsson A, Bjornsdóttir G, Jonsson JJ, Thorarensen O, Ludvigsson P, Gudbjartsson H, Eyjolfsson GI, Sigurdardóttir O, Olafsson I, Arnar DO, Magnusson OT, Kong A, Masson G, Thorsteinsdóttir U, Helgason A, Sulem P, Stefansson K (2015) Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* 47:435–444
 41. 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 Methods* 7:111–118
 42. Alsters SI, Goldstone AP, Buxton JL, Zekavati A, Sosinsky A, Yiorkas AM, Holder S, Klaber RE, Bridges N, van Haelst MM, le Roux CW, Walley AJ, Walters RG, Mueller M, Blakemore AI (2015) Truncating homozygous mutation of carboxypeptidase E (CPE) in a morbidly obese female with type 2 diabetes mellitus, intellectual disability and hypogonadotrophic hypogonadism. *PLoS One* 10:e0131417
 43. Stenson PD, Ball EV, Howells K, Phillips AD, Mort M, Cooper DN (2009) The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum Genomics* 4:69–72
 44. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 42:30–35
 45. Stranneheim H, Wedell A (2016) Exome and genome sequencing: a revolution for the discovery and diagnosis of monogenic disorders. *J Intern Med* 279:3–15
 46. Gussak I, Brugada P, Brugada J, Wright RS, Kopecky SL, Chaitman BR, Bjerregaard P (2000) Idiopathic short QT interval: a new clinical syndrome? *Cardiology* 94:99–102
 47. Gawad C, Koh W, Quake SR (2016) Single-cell genome sequencing: current state of the science. *Nat Rev Genet* 17:175–188
 48. Dean FB, Nelson JR, Giesler TL, Lasken RS (2001) Rapid amplification of plasmid and

- phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* 11:1095–1099
49. Evrony GD, Lee E, Mehta BK, Benjamini Y, Johnson RM, Cai X, Yang L, Haseley P, Lehmann HS, Park PJ, Walsh CA (2015) Cell lineage analysis in human brain using endogenous retroelements. *Neuron* 85:49–59
 50. Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, Lee S, Chittenden TW, D’Gama AM, Cai X, Luquette LJ, Lee E, Park PJ, Walsh CA (2015) Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* 350:94–98
 51. Leventer RJ, Guerrini R, Dobyns WB (2008) Malformations of cortical development and epilepsy. *Dialogues Clin Neurosci* 10:47–62
 52. Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. *Genetics* 148:1667–1686
 53. Biesecker LG, Spinner NB (2013) A genomic view of mosaicism and human disease. *Nat Rev Genet* 14:307–320
 54. Poduri A, Evrony GD, Cai X, Walsh CA (2013) Somatic mutation, genomic variation, and neurological disease. *Science* 341:1237–1241
 55. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. *Nature* 458:719–724
 56. Lohmann K, Klein C (2014) Next generation sequencing and the future of genetic diagnosis. *Neurotherapeutics* 11:699–707
 57. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073
 58. Dorschner MO, Amendola LM, Turner EH, Robertson PD, Shirts BH, Gallego CJ, Bennett RL, Jones KL, Tokita MJ, Bennett JT, Kim JH, Rosenthal EA, Kim DS, L. National Heart, P. Blood Institute Grand Opportunity Exome Sequencing, Tabor HK, Bamshad MJ, Motulsky AG, Scott CR, Pritchard CC, Walsh T, Burke W, Raskind WH, Byers P, Hisama FM, Nickerson DA, Jarvik GP (2013) Actionable, pathogenic incidental findings in 1,000 participants’ exomes. *Am J Hum Genet* 93:631–640
 59. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. *Science* 322:881–888
 60. Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 30:3894–3900
 61. Wang QF, Prabhakar S, Wang Q, Moses AM, Chanan S, Brown M, Eisen MB, Cheng JF, Rubin EM, Boffelli D (2006) Primate-specific evolution of an LDLR enhancer. *Genome Biol* 7:R68
 62. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444:499–502

Chapter 2

Induced Pluripotent Stem Cells in Disease Modeling and Gene Identification

Satish Kumar, John Blangero, and Joanne E. Curran

Abstract

Experimental modeling of human inherited disorders provides insight into the cellular and molecular mechanisms involved, and the underlying genetic component influencing, the disease phenotype. The breakthrough development of induced pluripotent stem cell (iPSC) technology represents a quantum leap in experimental modeling of human diseases, providing investigators with a self-renewing and, thus, unlimited source of pluripotent cells for targeted differentiation. In principle, the entire range of cell types found in the human body can be interrogated using an iPSC approach. Therefore, iPSC technology, and the increasingly refined abilities to differentiate iPSCs into disease-relevant target cells, has far-reaching implications for understanding disease pathophysiology, identifying disease-causing genes, and developing more precise therapeutics, including advances in regenerative medicine. In this chapter, we discuss the technological perspectives and recent developments in the application of patient-derived iPSC lines for human disease modeling and disease gene identification.

Key words Cellular reprogramming, iPSC, Human complex disease, Genetics

1 Introduction

Genetic linkage and genome-wide association studies (GWAS) have emerged as systematic approaches for identifying the root genetic causes of disease [1]. However, these approaches are often accompanied by certain challenges, such as: (1) independent genetic variants can produce similar phenotypes via molecularly distinct pathways, (2) a disease phenotype can emerge from the combined effect of multiple genetic factors, and (3) both linkage and association approaches focus on initial localization (albeit with differently sized support intervals) of the genetic loci. The challenge that investigators then face is how to mechanistically connect these genetic loci to the factors that initiate the disease process and ultimately lead to disease presentation. Because molecular pathways are shaped by cell-type-specific gene expression, it is preferable to model and study the molecular basis of a disease in the affected cell

or tissue type [2]. However, relevant human tissue or cell samples are often difficult to obtain, sometimes requiring invasive surgery or only becoming available post-mortem. Furthermore, isolated cells cannot be maintained or expanded with conventional culture conditions, and must instead be immortalized for long-term use. In the absence of primary tissue or cells of interest, animal models and heterologous or surrogate in vitro cell culture models have been invaluable tools for modeling human diseases. Transgenic models and gene targeting rely heavily on rodent models, and approximately 90% of animals used in research are mice, rats, and other rodents [3]. While rodents are, and will continue to be, extremely valuable models for biomedical research, rodents do not always accurately model human disease or biological response [4]. The evolutionary distance between rodents and humans (human-mouse-rat ancestor diverged ~87 million years ago) [5] presents significant differences in biological function that may limit the immediate translational value of findings. The close phylogenetic relationship and consequent similarity in biological processes and physiology of nonhuman primates to humans makes them better models for human diseases, but due to the difficult animal husbandry, cost, and ethical limitations, nonhuman primates account for only 0.28% of all the laboratory animals used in research [3]. Similarly, the human diseases and genetic disorder models utilizing patient-derived immortalized cell lines originating from surrogate tissues (i.e., blood or tissue biopsies) or utilizing cell types of interest from a heterologous species/system, often lacks the ability to faithfully recapitulate specific properties of the primary tissue of interest [6–10].

An alternative to these models is the stem cell-based system, which carries the intrinsic capability for indefinite self-renewal and the potential to model the tissue specific physiology through the use of differentiation protocols to generate specific target cell/tissue types. These properties enable us to study genotype–phenotype relationships in a broad range of human cell types and differentiation states, as well as obtain large numbers of cells for additional purposes, including drug screening and cell therapy [11]. Embryonic stem cell (ESC) lines were first established in mouse [12], and subsequently in human from in vitro derived embryos [13]. However, the challenges related to bioethics, safety, and the limited availability of disease-specific human embryonic stem cell (hESC) lines have complicated the utilization of this approach to its full potential. This changed dramatically in 2006 when Takahashi and Yamanaka made the seminal discovery that mouse skin fibroblasts, using a simple cocktail of pluripotency transcription factors, can be reprogrammed into an induced pluripotent stem cell (iPSC) state that shares the indefinite self-renewal and pluripotent differentiation capacities of ESCs [14]. One year later, these same investigators, as well as groups headed by James

Thomson and George Daley, succeeded in converting human fibroblasts to iPSCs [15–17]. Reprogramming to pluripotency has now been demonstrated starting with a variety of somatic cell types, including immortal cell lines [18–26]. The greatest advantage of iPSC technology is that it allows for the generation of pluripotent cells from any individual in the context of his or her own genetic identity. The technology has already been utilized in modeling sporadic, as well as complex, multifactorial diseases of unknown genetic identity by generating disease-specific cell types from patient somatic cells [27–38].

In this chapter, we describe the ways in which human iPSCs are generated and utilized for disease modeling, as well as for disease gene identification. We discuss common challenges and approaches that we and others have encountered in iPSC reprogramming, their differentiation into target cell types, and identification and measurement of disease relevant phenotypes, to provide an informed perspective of existing technologies and in vitro iPSC-based disease modeling in understanding human disease genetics.

2 iPSC Reprogramming

First, let us introduce an exemplary reprogramming method we use in our laboratory to reprogram human lymphoblastoid cell lines (LCLs) into iPSCs. The LCLs collected in genetic and epidemiological studies represents one of the largest, well-characterized, existing bioresources available for iPSC reprogramming, because a multitude of data already exists on sample donors. For example, the NIMH Repository and Genomic Resource alone currently stores over 184,000 LCLs for sharing with investigators of mental disorders [39].

To reprogram LCLs into iPSCs, the LCLs are propagated and while still in log growth phase, nucleofected with episomal plasmids (pCE-hOCT3/4, pCE-hSK, pCE-hUL, and pCE-mp53DD), encoding reprogramming factors (i.e., *OCT3/4*, *SOX2*, *KLF4*, *LMYC*, and *LIN28*), and mouse p53 carboxy-terminal dominant-negative fragment using the SE Cell Line 4D-Nucleofector X Kit and 4D-Nucleofector DN-100 program on a 4D-Nucleofector system (Lonza; <http://www.lonza.com/>). The plasmids are described in Okita et al. [40] and can be obtained from the Addgene plasmid repository. The nucleofected LCLs are allowed to recover for 8–12 h in LCL growth media (RPMI 1640 complete media; all media components from Life Technologies) in a CO₂ incubator at 37 °C, 5% CO₂ and atmospheric O₂, and then transferred onto a Matrigel matrix (Corning Inc.)-coated, six-well plate and cultured in iPSC reprogramming media (TeSR-E7 from STEMCELL Technologies) for 12–14 days. From days 13–15, when iPSC-like colonies (Fig. 1a) start to appear, the cultures are

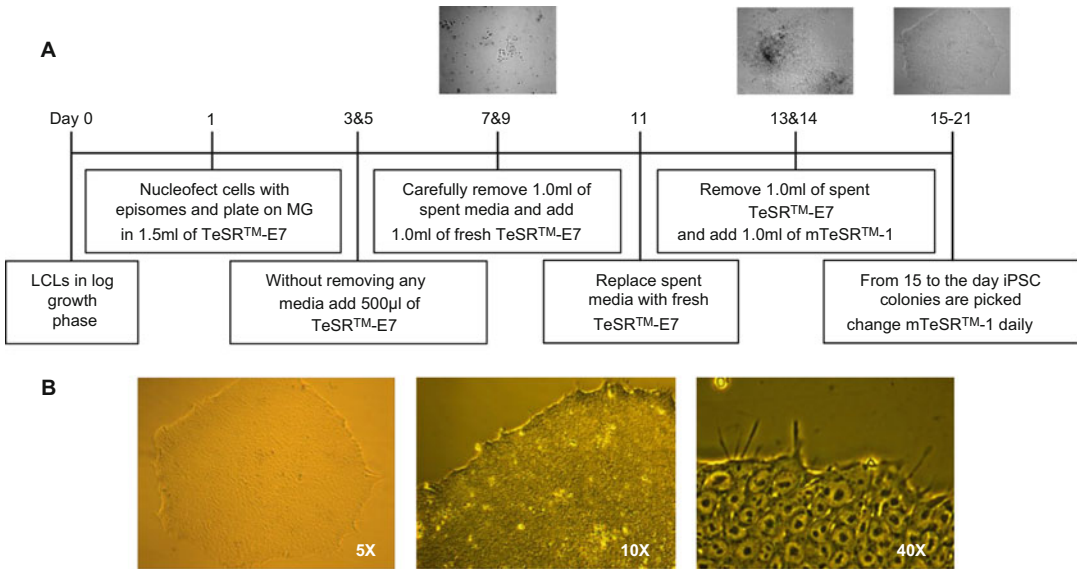


Fig. 1 LCL to iPSC reprogramming. **(a)** Schematic diagram of LCL to iPSC reprogramming. **(b)** Morphology of a reprogrammed iPSC colony at 5×, 10×, and 40× original magnifications, respectively

transitioned to human iPSC/ESC maintenance media (mTeSR-1 from STEMCELL Technologies). Media is changed daily thereafter. On days 18–21, 10 to 15 colonies morphologically similar to human ESCs (Fig. 1b) are manually picked for further cultivation and evaluation of stem cell characteristics, such as expression of pluripotency markers, differentiation potential, and genomic integrity. Further details on this LCL-to-iPSC reprogramming protocol can be found in Kumar et al. [26]. Using this protocol, we achieved high reprogramming efficiency and a 100% success rate. However, the reprogramming efficiency can vary significantly among cell lines, based on our unpublished findings in more than 50 iPSC lines. Also, the differential gene expression analysis of the cellular and EBV viral genes, as well as the quantitative PCR analysis of the EBV DNA in the LCL reprogrammed iPSCs, shows that transcription and replication of the EBV genome are inhibited in the reprogrammed iPSCs, which ultimately results in the complete depletion of the EBV genome from the reprogrammed iPSCs [26, 41, 42].

The common aim of all somatic cell reprogramming methods is the forced expression of reprogramming factors into the cells to be reprogrammed. However, depending on what type of cells are being reprogrammed, the capacity and efficiency with which the reprogramming method can deliver reprogramming factors into the cell types used and the downstream applications of the generated iPSCs, certain technologies have advantages over the others.

Therefore, to better understand the reprogramming process and provide a basis for selecting the most suitable method, we present a comprehensive overview of the most popular reprogramming approaches.

2.1 Retroviral and Lentiviral Reprogramming

Retroviruses, a family of viruses that stably integrate into the host genome, are one of the most common types of viruses used to express genes at robust levels in mammalian cells, and have been used for several decades. The first iPSC reprogramming studies utilized retroviral vectors to express reprogramming factors [14, 15]. While retroviruses generally only infect dividing cells, because their access to the host genome is thought to rely on the breakdown of the nuclear envelope that occurs during mitosis [43], lentiviruses are a genus of the retroviral family that can infect non-dividing cells, possibly through the use of nuclear localization signals by the viral components [44]. In an attempt to improve reprogramming efficiency by utilizing both dividing and nondividing cells, the Thomson lab was the first to successfully demonstrate the use of lentivirus in iPSC reprogramming [16]. Though the retro and lentiviral vectors proved to be robust delivery vehicles for iPSC reprogramming, the major drawback of these delivery systems was the random integration of the viral genome carrying transgenes into the cellular genome. For example, the integration site may be a regulatory or a structural element [45]. Secondly, the copy number of viral genomes integrated into the cellular genome may vary to a great extent from experiment to experiment [15]. The presence of the transgenes in the reprogrammed iPSC may not only hinder their clinical use, but may be of concern in some disease modeling and gene discovery strategies [46].

The first generation of transgene-free iPSCs was generated with lentiviral vectors containing loxP sites in the 5' and 3' LTR of the viral vectors. The presence of loxP sites provided a substrate to remove most of the transgene sequences by Cre-mediated recombination [47–50]. However, this strategy removed almost all of the transgene, except one loxP site flanked by small portions of the 5' and 3' LTRs, which remains in the iPSC genome. Also, this reprogramming strategy requires an additional step after iPSC reprogramming for the excision of the transgene, such that only a small portion of the reprogramming vector remains integrated in the iPSC genome. Another consideration would be how amenable the starting material (i.e., the cells to be reprogrammed) is to transduction. For example, LCLs show a poor transduction efficiency (~0.8%) compared to human skin fibroblasts (~88%) [42].

2.2 Nonintegrating Reprogramming Methods

There are several reprogramming methods that leave no trace of transgenes in the reprogrammed iPSCs (i.e., Sendai virus or adenovirus, episomal plasmids or minicircles, direct transfection with reprogramming mRNA, miRNA, or protein, and transposition

with the piggyBac transposon). However, some of these nonintegrating reprogramming methods either have poor reprogramming efficiency (adenovirus and proteins), or are ineffective in reprogramming any somatic cells other than fibroblasts (minicircles), or require post-reprogramming excision of the vector sequence (piggyBac). Because of these issues, we focus here on the non-integrating reprogramming methods that are widely practiced and use readily available reagents and kits.

2.2.1 Sendai Virus (SeV)

Since its isolation in the 1950s [51], SeV has occupied a unique position as a research tool for basic and applied biology. The SeV is an RNA virus that remains in the cytoplasm (i.e., there is no integration in the host genome) of the infected cells and can robustly express reprogramming genes for a few passages. However, SeV quickly becomes diluted out and eventually lost completely. Wild-type SeV vectors installed with Oct4, Sox2, Klf4, and c-Myc cDNA were reported to generate transgene-free iPSCs, dependent on passive elimination of the vector genome through cell passage [52]. This prototype was replaced with a less cytotoxic backbone [23] with a temperature-sensitive (ts) mutation that facilitated the faster clearance of the vector genome [53], and is now commercially available. The SeV reprogramming is efficient, highly reliable, and works with many different somatic cell types with a low workload and a complete absence of viral sequences in most lines at higher passages [54]. The shortcomings of SeV include relatively slow clearance of SeV RNA and the current lack of clinical-grade SeV for reprogramming.

2.2.2 Episomal Plasmids

In episomal plasmid-based reprogramming, prolonged expression of reprogramming factors is achieved by oriP/EBNA1-based episomal vectors. These plasmids contain Epstein-Barr virus derived oriP/EBNA1 viral elements, which facilitate episomal plasmid DNA replication in dividing cells [55–57], and thus allow expression of reprogramming factors for a long enough period to initiate the reprogramming process while eventually being lost from proliferating cells, leaving no footprint of the transfected plasmid. Human episomal reprogramming was first demonstrated by the Thomson laboratory, using a single transfection of three plasmids containing Oct4, Sox2, Nanog, and Klf4; Oct4, Sox2, and SV40 Large T antigen; and c-myc and Lin28 [58]. However, the reprogramming efficiencies were significantly low. Since the publication of the first successful episomal reprogramming, a concerted effort has been extended to improve the efficiency of this method. Utilizing the same set of episomal plasmids, Hu et al. [59] reprogrammed bone marrow- and cord blood-derived mononuclear cells and confirmed the low reprogramming efficiency observed in fibroblasts in the earlier study. However, they found the addition of “Thiazovivin,” a small molecule identified in a previous chemical

screen to improve hESC survival during passaging [60], enhanced the reprogramming efficiency by ten-fold. This study also provided a more detailed description of the plasmid loss, showing that all of the generated iPSC lines lost the plasmid between passages 3–15. Later in the same year, Yu et al. [61] published feeder-free conditions for the episomal reprogramming that included the improvement of reprogramming efficiencies using a cocktail containing MEK inhibitor PD0325901, GSK3b inhibitor CHIR99021, TGF- β /Activin/Nodal receptor inhibitor A-83-01, ROCK inhibitor HA-100, and human leukemia inhibitory factor. Now, reprogramming media that work with a wide variety of cell types are commercially available. Using an oriP/EBNA1 plasmid constructed with Yamanaka factors (Oct4, Sox2, Klf4, and c-Myc) plus Lin28 and another oriP/EBNA1 plasmids expressing SV40 large T antigen, and p53-shRNA, Chou et al. [62] demonstrated highly efficient reprogramming of briefly cultured blood mononuclear cells. Further improvements in the reprogramming efficiency were made in a method published by Okita et al. [63], which employs the reprogramming factors Oct4, Sox2, Klf4, L-Myc, and Lin28A, combined with p53 knockdown (p53-shRNA/mp53DD). While optimizing an efficient method for LCL to iPSC reprogramming, we used a similar strategy and confirmed previous findings [40, 63] that p53 knockdown and removal of SV40 large T antigen improved reprogramming efficiency and success considerably [26]. Key advantages of episomal plasmid-based reprogramming are the high reliability of iPSC generation from a variety of cell types (e.g., skin fibroblast, blood-derived CD34+ and peripheral blood mononuclear cells, and stored LCLs) [26, 40, 63] and the quick loss of the reprogramming agent (relative to SeV) [54]. However, episomal reprogramming may raise concerns regarding the genetic integrity of the resulting iPSC lines due to the use of p53 knockdown. Notably, we and others did not observe any significant increase in structural abnormalities in reprogrammed iPSCs [26, 54, 63].

2.2.3 mRNA Transfection

In mRNA reprogramming, cells are transfected with in vitro transcribed mRNAs that encode reprogramming factors; however, this strategy requires mitigating the strong immunogenic response elicited in cells due to the introduction of synthetic nucleic acid. Warren et al. [64], reported the first successful synthetic mRNA-based reprogramming, employing several measures to limit activation of the innate immune system by foreign nucleic acids. They modified RNA bases by substituting 5-methylcytidine for cytidine and pseudouridine for uridine, and added the interferon inhibitor B18R into cell culture media [64]. The main advantages of the RNA method are the speed of colony emergence, comparatively high reprogramming efficiency, a complete absence of integration,

a very low aneuploidy rate, and a low donor cell requirement (typically 50,000 cells, but as few as 1000 human fibroblasts can be reprogrammed) [54]. However, due to the very short half-life of mRNAs, daily transfections are required to induce iPSCs, which significantly increases hands-on time and overall workload. Also the protocol requires a tissue culture incubator with oxygen control [64]. More significantly, RNA reprogramming shows poor success rate with different samples. Studies have reported frequent unsuccessful attempts to reprogram patient-derived primary fibroblasts [54, 65]. Further, there have been no reports to date of successful generation of iPSCs from blood-derived cells using this methodology. One promising approach for improving mRNA-based reprogramming is the inclusion of pluripotency-inducing miRNAs. Warren et al. [66] reported increased RNA reprogramming efficiencies and accelerated colony emergence by fusing Oct4 to a heterologous transactivation domain. The conventional mRNA method can therefore be useful for easy-to-reprogram fibroblast samples, but needs to be further optimized to overcome the reprogramming resistance and excessive cell death observed with many patient samples.

3 iPSC-Based Human Disease Modeling

The greatest advantage of iPSC technology in *in vitro* disease modeling is that it allows for the generation of pluripotent cells from any individual in the context of his or her own genetic identity, including individuals with sporadic forms of disease, as well as those affected by complex multifactorial diseases of unknown genetic identity [29]. The generated iPSCs from patients and suitable controls are differentiated into target cell types affected in the given disease and compared for disease-relevant phenotypes [67]. Each stage in this process poses challenges. What are the appropriate cases and controls to include? How to identify and generate disease-relevant target cell types? How to deal with the heterogeneous mix of cell types that results from iPSC differentiation? How to identify and analyze cellular phenotypes relevant to the disease mechanism? Here, we discuss these challenges and present potential approaches for addressing each.

3.1 Cases and Controls

It is now well recognized that the genetics of iPSCs reflect the genetics of the patient; the vast majority of the transcriptional and epigenetic signatures and differentiation propensities are donor-determined [68–72]. These properties, on the one hand, make iPSCs a great tool to model human diseases, but on the other hand, render them notoriously variable in phenotypic output and differentiation propensities [73, 74]. There can be a possibility of missing phenotypic effect in the phenotypic noise caused by the

variable genetic backgrounds of unrelated iPSC lines [67]. Therefore, because of the inherent differences resulting from the donor-dependent variability, it seems obvious that relatively large cohorts of iPSC lines from different donors, representing both cases and controls, would be needed to obtain reliable results concerning the impact of donor-specific variants. This approach has been successfully applied using a relatively small sample size of 4–14 different patient-specific iPSC lines compared to similar numbers of control cells for the identification of disease-specific cellular phenotypes [36, 38, 75]. The derivation of iPSCs from multiple patients, though labor intensive, is usually straightforward, enabling the analysis of similar mutations in diverse genetic backgrounds. In addition, patient-derived iPSCs are more beneficial than genome editing in normal iPSCs when modeling genetically complex disorders, which often involve multiple unknown loci [11]. It is thus understandable that national and international initiatives are already investing in major efforts to establish repositories of human iPSCs as models for human disorders. These repositories are aimed at generating thousands of new cell lines, for both monogenic and complex disorders, and using nonintegrative reprogramming methods such as the use of Sendai viruses or episomal vectors [11]. Also repurposing the LCL repositories, generated and maintained in genetic and epidemiological studies worldwide, holds great potential for generating iPSCs to model human diseases particularly for disease gene identification [26].

3.2 Generating Disease-Relevant Target Cell Types

The first challenge when modeling a disease *in vitro* is identifying the target cell type to investigate. Broadly speaking, iPSC-based disease modeling is more relevant in cases where studies of patient tissues have identified the cell types whose loss or dysfunction causes the disease; however, target tissue and cells are difficult to obtain or cannot be maintained or expanded with conventional culture conditions. The repertoire of cell types that can be generated *in vitro* from iPSCs is impressive, but still small compared to the number of cell types in the human body. Cell types affected in disease are generated from iPSCs by directed differentiation. The directed differentiation protocols utilize signaling pathways that are responsible for *in vivo* differentiation of the target cell types. The signaling pathways, are stimulated or inhibited *in vitro* by biological (recombinant growth factors) or small-molecule modulators added at specific times and concentrations [67, 76–79]. Although the differentiation efficiency and quality of generated target cell-type are constantly improving; the process is often inefficient and produces a heterogeneous cell population consisting of multiple cell types or a mixture of cells at different developmental stages, mostly consisting of fetal or immature phenotypes (for example, cardiomyocytes [80, 81], dendritic cells [82], neural cells [79], and pancreatic β -cells [83]). Because restricting phenotypic and

molecular analyses to a relatively homogeneous target cell population would facilitate comparison across different cell lines, it is necessary to characterize and purify the disease-relevant target cells. Some cellular phenotypes, including survival, morphology, and protein expression or localization, can be identified by immunostaining analysis of target cell-specific markers and candidate disease proteins. For a wider array of experimental manipulations and analyses, target cell populations can be purified by unique combinations of surface markers, genetically encoded reporter genes, or drug-resistance genes [67, 84–86]. Enrichment and maturation protocols that modulate culture media components may also be utilized to achieve a mature phenotype in the target cell population [81]. In addition, many cases will require iPSC line-specific optimization and modifications for efficient differentiation into target cells [78]. It is remarkable that, despite these challenges, target cells derived in vitro by iPSC differentiation often display phenotypes observed in their mature counterparts in vivo. For example, cellular phenotypes have been seen in models of late-onset neurodegenerative diseases such as Parkinson’s disease, schizophrenia, and Alzheimer’s disease [36, 38, 87–89]. Also, cardiac disease phenotypes, such as cardiac hypertrophy, can be modeled into iPSC-derived cardiomyocytes following a simple maturation step [81].

An alternative to the iPSC-based, directed differentiation approach is “direct programming,” which relies on forced gene expression, generally of relevant transcription factors or microRNAs, for converting one cell type into another resembling the target cells [90–94]. While this approach is promising, it is still unclear to what extent these programmed cells are suitable for in-vitro disease modeling, because they may be less like their in-vivo counterparts than cells generated by directed differentiation of iPSCs [67, 95].

3.3 Identification and Analysis of Disease-Relevant Cellular Phenotypes

For disease gene identification and validation, the goal of in vitro disease modeling is to unveil poorly understood or unknown disease mechanism(s) and relate them to the underlying genetic component. The key challenge to this process is to identify relevant cellular phenotypes that accurately represent the disease pathophysiology and bridge the gap to causal genetic mechanisms. Increasing numbers of reports have demonstrated that for many diseases, the specific pathophysiology can be captured in human iPSC-based disease models. These range from cardiovascular disease [37, 81, 96], cancer [97, 98], ocular disease [99, 100], diabetes mellitus [101, 102], and neurological disorders of the brain [103, 104]. Similarly, approaches utilizing phenotypes that are sensitive, unbiased, and measurable at genome-wide scales might be most relevant to the discovery of molecular changes and downstream candidate gene or genetic variants. Post-genomic technologies offer a battery

of approaches for profiling cell differences at both the population and single-cell levels. Advances in RNA-sequencing technologies and transcriptomics provide one of the easiest and highest throughput approaches to cell phenotyping, and have been traditionally used for disease gene identifications using primary or surrogate in vitro cell models [105, 106]. Transcriptome studies of both schizophrenia and autism spectrum disorder patient-iPSC derived cells have identified hundreds of gene expression differences [35, 38, 107–109]. Mapping and measuring DNA methylation and chromatin accessibility (ATAC Seq) may extend this analysis to provide unique epigenetic signatures, as seen for example with the methyl-cytosine-binding protein MeCP2 that causes Rett's syndrome and is associated with autism spectrum disorders [110]. Histone protein modifications can be profiled using ChIP-seq and several histone methyl transferase enzymes are associated with neuropsychiatric disorders. Either alone, or more likely when combined with expression data, epigenetic profiling may identify developmental and activity-dependent cellular phenotypes [111–113]. Proteomic technologies could, in turn, be used to back up the results of transcriptional profiling by measuring the quantitative changes in protein levels [38] and identifying the specific binding partners of candidate protein(s) in target cell types [114, 115].

Another important consideration of cellular phenotyping is the differentiation stage and the time point at which the target cells are assayed. There are two issues, one is the maturation state of the target cells, which we have discussed briefly in the previous section, and the second is the expected time course of a given disease process. For congenital or early-onset diseases, it may be sufficient to model the disease in immature cells at early time points in vitro. For late-onset diseases, it is less clear when the disease process begins. For example, characteristic symptoms of schizophrenia generally appear late in adolescence, and the disease is thought to be a neurodevelopmental condition [116] that is often predated by a prodromal period that can appear in childhood [117]. Because damaging de novo mutations in persons with schizophrenia converge in a network of genes coexpressed in the prefrontal cortex during fetal development, one prevailing hypothesis is that disruptions in fetal prefrontal cortical development underlie schizophrenia [118]. In human iPSC-based in vitro models of schizophrenia, although iPSC derived neurons are electrophysiologically active, gene expression patterns indicate that they are immature relative to those in the human brain [119, 120]. iPSCs can be differentiated into cortical pyramidal [121] and interneuron fates [120, 122], but these neurons require months to fully mature in vitro and generally lack myelination [123, 124]. Neural progenitor cells (NPCs) are a highly replicative neural population capable of rapidly initiating neuronal differentiation; they are easily assayed, well-suited for

scalability, and reveal reproducible schizophrenia associated transcriptomic/gene expression phenotypes [38]. Conversely, iPSC-derived human cardiomyocytes only display disease-associated phenotypes in an adult-like state [125]. One approach to address this issue is to artificially “age” target cells by challenging them with an environmental stressor. This approach revealed a selective sensitivity in disease-derived dopaminergic neurons that otherwise appeared indistinguishable from controls [87, 89].

4 Approaches to Identify and Validate Disease Genes

In recent years, genome-wide association studies (GWAS) have successfully tagged thousands of disease- or trait-associated genetic loci. However, molecular mechanisms linking genetic loci to a disease phenotype often remain unclear. Moreover, for most complex diseases and traits, associations found in GWAS explain only a small proportion of the phenotypic variation [126, 127]. For example, although 71 independent loci have been associated with Crohn’s disease, they account for only 23% of the estimated heritability [128]. GWAS of psychiatric diseases show an even less favorable picture. For instance, schizophrenia has an estimated heritability of 80% [129, 130], but observed genetic variants currently account for <1% of the variance [131]. To bridge this gap between genotype and disease phenotype and to better understand the biological mechanisms and translational possibilities, the genotype-driven approach, aided with deep phenotyping, appears to be a more appealing and powerful strategy [132, 133]. A limitation of this approach however, is that the disease pathologies are often tissue- or cell type-specific [134–138], and due to ethical and practical reasons, deep phenotyping analyses are often only feasible in easily available surrogate tissues such as blood, particularly in large population-based gene identification studies. The iPSC technologies discussed here in this chapter, along with new sequencing technologies, genome-wide assays, and comprehensive genome annotation are now offering opportunities to interrogate genome function in multiple individuals at the cellular and tissue level. Together with more precise characterization of clinical and pathophysiologic phenotypes, a range of deep cell- or tissue-specific phenotypes (sometimes referred to as “endophenotypes” or intermediate phenotypes) and “omic-metrics” such as epigenomics, transcriptomics, proteomics, and metabolomics can be examined in an integrated fashion in disease-relevant cells or tissues to understand the molecular mechanisms underlying the phenotypic expression of diseases [139]. It is hoped that ongoing improvements in iPSC reprogramming, differentiation, and disease modeling capabilities will facilitate the analysis of more functionally specific

phenotypes in large-scale study samples that will exhibit higher genetic signal-to-noise ratios and speed causal gene identification.

iPSC-based disease modeling also opened new avenues for developing faster and more reliable assays to investigate and validate the biological context of the genetically identified disease loci. The iPSC lines can be reprogrammed from cells (e.g., skin fibroblast or blood-derived cells) of individual(s) carrying the genotype of interest, representing an index variant or other potential functional variant(s) in linkage disequilibrium (LD) with the index variant. The generated iPSC lines are then genome-edited to correct for the disease- or trait-associated risk allele (risk allele to wild type) and differentiated into target cell types. This will generate isogenic control cells that differ from the original disease/patient specific target cells only at the target variant(s). These corrected target cells are the perfect control for a comparative analysis using genome-wide “omic-metrics” to understand the molecular and biological context of the risk variant(s). Nuclease-based genome editing techniques have seen great improvements in recent years, and several technologies allowing targeted manipulation of the human genome using designer proteins or protein-RNA hybrids that recognize specific DNA sequences, exist to perform such genetic manipulation/correction [140]. These tools include zinc fingers (ZFs), transcription activator-like effectors (TALEs), and the CRISPR-Cas9 system [141–145]. ZFs and TALEs are DNA-binding proteins that can be fused to nucleases such as FokI to generate ZFNs and TALENs [67, 146, 147]. FokI acts as an obligate dimer, and DNA double-strand breaks (DSBs) are only generated when FokI monomers are brought together by ZFNs or TALENs targeting adjacent DNA sequences. The bacterial CRISPR-Cas9 system, on the other hand, uses a combination of proteins and short guide RNAs to recognize and cleave complementary DNA sequences via a nuclease (Cas9) [148]. When ZFNs, TALENs, or Cas9-guide RNAs are transfected into human iPSCs, along with a targeting construct containing homology arms 5' and 3' to the induced DSB site, the lesion can be repaired by homologous recombination (HR), which inserts the targeting construct into the genomic region of interest [67, 149]. There are several reports demonstrating the feasibility of performing genome editing in human pluripotent stem cells (PSCs) and iPSCs with ZFNs, TALENs, CRISPRs, and other tools [145, 150–158]. A few studies, mostly on well characterized, disease-causing variants, have used genome-editing tools to generate isogenic wild-type versus mutant cell lines that have then been differentiated into disease-relevant cell types and shown to display phenotypic differences that give insight into disease pathophysiology. For example, Reinhardt et al. [89] generated iPSCs from Parkinson's disease (PD) patients carrying a G2019S mutation (rs34637584) in the *LRRK2* gene. This mutation is associated with familial and sporadic PD. They used ZFNs to

correct the mutation in three of the patient-derived lines and insert the mutation into a control iPSC line. The matched cell lines were then differentiated into midbrain dopaminergic (mDA) neurons. Expression profiling of pairs of isogenic wild-type and mutant mDA cell lines revealed several genes that are consistently dysregulated by the mutant *LRRK2* gene, including *CPNE8*, *CADPS2*, *MAP7*, and *UHRF2*; remarkably, individual knockdown of each of those genes in mutant neurons modulated their sensitivity to oxidative stress. The investigators also established that the increased sensitivity to stress of the mutant neurons was at least in part due to activation of ERK signaling and could be reversed with an inhibitor of ERK phosphorylation.

5 Conclusion and Future Perspectives

The proof-of-concept emerging from many recent studies that have attempted to model complex disease and disorders in vitro using iPSC-derived patient target cells has been promising. For example, iPSC-derived neurons and neural progenitor cells (NPCs) to model schizophrenia and Alzheimer's disease [35, 36, 38], iPSC-derived hepatocytes to model inherited metabolic disorders of the liver [31], and iPSC-derived cardiomyocytes to model hypertrophic cardiomyopathies and diabetes-induced cardiomyopathies [81, 159], have been very encouraging. Improved methodologies allowing reliable iPSC reprogramming from more easily accessible cells such as blood-derived cells and stored LCLs, improved and simplified target cell differentiation protocols, and an integrative genome-wide approach to identify and develop standardized cellular phenotypes that accurately represent disease pathophysiology and bridge the gap to causal genetic mechanisms, will provide the solution to many problems. However, there remain a number of considerable challenges ahead.

In the future, strategies will need to consider the genetic characteristics of complex diseases and disorders, where despite the very high measurable genetic component, identification of the causal genes and variants has proved daunting. In complex diseases and disorders, genetic risk is largely polygenic, with a mixture of many common variants of small effect, as well as few rare variants of large effect. In contrast, a priori we would expect to find the most robust phenotypes in cells derived from patients carrying highly penetrant rare variants. It will be important to connect the knowledge gained from single gene deficits with that gained from the accumulated effects of multiple subtle genetic risk alleles.

Both the selection of patients carrying rare variants of large effect and the selection of patients of extremely high polygenic risk require large patient populations to optimize the selection. When genetic risk in selected patients is not sufficiently causal, any iPSC

experiment will require the analysis of large numbers of patient cell lines. This will require standardization and rigorous quality control to reduce technical variation to an acceptable minimum. Given the currently high reagent costs and labor-intensive nature of stem cell research, considerable improvements and alternative strategies are needed to integrate these processes with global efforts in patient recruitment and accompanying clinical phenotyping and genomic analysis. The LCL repositories that exist in many large-scale genetic studies, where a multitude of data, including whole genome DNA sequences of donors, provides an excellent opportunity to integrate and utilize this technology in gene identification.

Finally, beyond the issues of variability and capacity in generating target cells lies the key question of what are the relevant cellular phenotype(s) that can be typed efficiently in large sample sizes? We have discussed this somewhat in detail in the previous section. We argue that genome-wide tools, such as whole genome gene expression using microarray or RNAseq, might be most relevant in disease gene identification. Quantitative differences in gene expression can be directly correlated to the presence or absence of the disease or other disease-relevant phenotypes. This approach has already been applied in disease gene identification with mixed success using surrogate cell models. We believe, due to the tissue-specific variation in gene expression architecture, whole genome gene expression data generated from iPSC-derived, disease-relevant target cells, coupled with whole genome sequence data and other “omic metric” data, will provide a more integrated platform for disease gene identification.

References

1. Lander ES (2011) Initial impact of the sequencing of the human genome. *Nature* 470(7333):187–197. <https://doi.org/10.1038/nature09792>
2. Handley A, Schauer T, Ladurner AG et al (2015) Designing cell-type-specific genome-wide experiments. *Mol Cell* 58(4):621–631. <https://doi.org/10.1016/j.molcel.2015.04.024>
3. Phillips KA, Bales KL, Capitanio JP et al (2014) Why primate models matter. *Am J Primatol* 76(9):801–827. <https://doi.org/10.1002/ajp.22281>
4. Seok J, Warren HS, Cuenca AG et al (2013) Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc Natl Acad Sci U S A* 110(9):3507–3512. <https://doi.org/10.1073/pnas.1222878110>
5. Springer MS, Murphy WJ, Eizirik E et al (2003) Placental mammal diversification and the cretaceous-tertiary boundary. *Proc Natl Acad Sci U S A* 100(3):1056–1061. <https://doi.org/10.1073/pnas.0334222100>
6. Masters JR, Stacey GN (2007) Changing medium and passaging cell lines. *Nat Protoc* 2(9):2276–2284
7. Min JL, Barrett A, Watts T et al (2010) Variability of gene expression profiles in human blood and lymphoblastoid cell lines. *BMC Genomics* 11:96. <https://doi.org/10.1186/1471-2164-11-96>
8. Caliskan M, Cusanovich DA, Ober C et al (2011) The effects of EBV transformation on gene expression levels and methylation profiles. *Hum Mol Genet* 20(8):1643–1652. <https://doi.org/10.1093/hmg/ddr041>
9. Nestor CE, Ottaviano R, Reinhardt D et al (2015) Rapid reprogramming of epigenetic and transcriptional profiles in mammalian culture systems. *Genome Biol* 16:11. <https://doi.org/10.1186/s13059-014-0576-y>

10. Horvath P, Aulner N, Bickle M et al (2016) Screening out irrelevant cell-based models of disease. *Nat Rev Drug Discov* 15 (11):751–769. <https://doi.org/10.1038/nrd.2016.175>
11. Avior Y, Sagi I, Benvenisty N (2016) Pluripotent stem cells in disease modelling and drug discovery. *Nat Rev Mol Cell Biol* 17 (3):170–182. <https://doi.org/10.1038/nrm.2015.27>
12. Evans MJ, Kaufman MH (1981) Establishment in culture of pluripotential cells from mouse embryos. *Nature* 292(5819):154–156
13. Thomson JA, Itskovitz-Eldor J, Shapiro SS et al (1998) Embryonic stem cell lines derived from human blastocysts. *Science* 282 (5391):1145–1147
14. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126(4):663–676
15. Takahashi K, Tanabe K, Ohnuki M et al (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131(5):861–872
16. Yu J, Vodyanik MA, Smuga-Otto K et al (2007) Induced pluripotent stem cell lines derived from human somatic cells. *Science* 318(5858):1917–1920
17. Park IH, Zhao R, West JA et al (2008) Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* 451 (7175):141–146
18. Aasen T, Raya A, Barrero MJ et al (2008) Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. *Nat Biotechnol* 26(11):1276–1284. <https://doi.org/10.1038/nbt.1503>
19. Hanna J, Markoulaki S, Schorderet P et al (2008) Direct reprogramming of terminally differentiated mature B lymphocytes to pluripotency. *Cell* 133(2):250–264. <https://doi.org/10.1016/j.cell.2008.03.028>
20. Utikal J, Maherali N, Kulalert W et al (2009) Sox2 is dispensable for the reprogramming of melanocytes and melanoma cells into induced pluripotent stem cells. *J Cell Sci* 122 (Pt 19):3502–3510. <https://doi.org/10.1242/jcs.054783>
21. Carette JE, Pruszk J, Varadarajan M et al (2010) Generation of iPSCs from cultured human malignant cells. *Blood* 115 (20):4039–4042. <https://doi.org/10.1182/blood-2009-07-231845>
22. Miyoshi N, Ishii H, Nagai K et al (2010) Defined factors induce reprogramming of gastrointestinal cancer cells. *Proc Natl Acad Sci U S A* 107(1):40–45. <https://doi.org/10.1073/pnas.0912407107>
23. Seki T, Yuasa S, Oda M et al (2010) Generation of induced pluripotent stem cells from human terminally differentiated circulating T cells. *Cell Stem Cell* 7(1):11–14. <https://doi.org/10.1016/j.stem.2010.06.003>
24. Tsai SY, Clavel C, Kim S et al (2010) Oct4 and klf4 reprogram dermal papilla cells into induced pluripotent stem cells. *Stem Cells* 28(2):221–228. <https://doi.org/10.1002/stem.281>
25. Kim J, Lengner CJ, Kirak O et al (2011) Reprogramming of postnatal neurons into induced pluripotent stem cells by defined factors. *Stem Cells* 29(6):992–1000. <https://doi.org/10.1002/stem.641>
26. Kumar S, Curran JE, Glahn DC et al (2016) Utility of lymphoblastoid cell lines for induced pluripotent stem cell generation. *Stem Cells Int* 2016:2349261. <https://doi.org/10.1155/2016/2349261>
27. Rubin LL (2008) Stem cells and drug discovery: the beginning of a new era? *Cell* 132 (4):549–552. <https://doi.org/10.1016/j.cell.2008.02.010>
28. Maehr R, Chen S, Snitow M et al (2009) Generation of pluripotent stem cells from patients with type 1 diabetes. *Proc Natl Acad Sci U S A* 106(37):15768–15773. <https://doi.org/10.1073/pnas.0906894106>
29. Chun YS, Chaudhari P, Jang YY (2010) Applications of patient-specific induced pluripotent stem cells; focused on disease modeling, drug screening and therapeutic potentials for liver disease. *Int J Biol Sci* 6(7):796–805
30. Ghodsizadeh A, Taei A, Totonchi M et al (2010) Generation of liver disease-specific induced pluripotent stem cells along with efficient differentiation to functional hepatocyte-like cells. *Stem Cell Rev* 6(4):622–632. <https://doi.org/10.1007/s12015-010-9189-3>
31. Rashid ST, Corbineau S, Hannan N et al (2010) Modeling inherited metabolic disorders of the liver using human induced pluripotent stem cells. *J Clin Invest* 120 (9):3127–3136. <https://doi.org/10.1172/JCI43122>
32. Rosenzweig A (2010) Illuminating the potential of pluripotent stem cells. *N Engl J Med* 363(15):1471–1472. <https://doi.org/10.1056/NEJMe1007902>
33. Yoshida Y, Yamanaka S (2010) Recent stem cell advances: induced pluripotent stem cells for disease modeling and stem cell-based regeneration. *Circulation* 122(1):80–87.

- <https://doi.org/10.1161/CIRCULATIONAHA.109.881433>
34. Zhang N, An MC, Montoro D et al (2010) Characterization of human Huntington's disease cell model from induced pluripotent stem cells. *PLoS Curr* 2:RRN1193. <https://doi.org/10.1371/currents.RRN1193>
 35. Brennand KJ, Simone A, Jou J et al (2011) Modelling schizophrenia using human induced pluripotent stem cells. *Nature* 473(7346):221–225. <https://doi.org/10.1038/nature09915>
 36. Kondo T, Asai M, Tsukita K et al (2013) Modeling Alzheimer's disease with iPSCs reveals stress phenotypes associated with intracellular Abeta and differential drug responsiveness. *Cell Stem Cell* 12(4):487–496. <https://doi.org/10.1016/j.stem.2013.01.009>
 37. Liang P, Du J (2014) Human induced pluripotent stem cell for modeling cardiovascular diseases. *Regen Med Res* 2(1):4. <https://doi.org/10.1186/2050-490X-2-4>
 38. Brennand K, Savas JN, Kim Y et al (2015) Phenotypic differences in hiPSC NPCs derived from patients with schizophrenia. *Mol Psychiatry* 20(3):361–368. <https://doi.org/10.1038/mp.2014.22>
 39. NIMH-RGR Data Explorer (2015) NIMH Repository and Genomics Resource, USA. <https://explorer.nimhgenetics.org/>. Accessed 14 Oct 2015
 40. Okita K, Yamakawa T, Matsumura Y et al (2013) An efficient nonviral method to generate integration-free human-induced pluripotent stem cells from cord blood and peripheral blood cells. *Stem Cells* 31(3):458–466. <https://doi.org/10.1002/stem.1293>
 41. Rajesh D, Dickerson SJ, Yu J et al (2011) Human lymphoblastoid B-cell lines reprogrammed to EBV-free induced pluripotent stem cells. *Blood* 118(7):1797–1800. <https://doi.org/10.1182/blood-2011-01-332064>
 42. Choi SM, Liu H, Chaudhari P et al (2011) Reprogramming of EBV-immortalized B-lymphocyte cell lines into induced pluripotent stem cells. *Blood* 118(7):1801–1805. <https://doi.org/10.1182/blood-2011-03-340620>
 43. Roe T, Reynolds TC, Yu G et al (1993) Integration of murine leukemia virus DNA depends on mitosis. *EMBO J* 12(5):2099–2108
 44. Bukrinsky MI, Sharova N, Dempsey MP et al (1992) Active nuclear import of human immunodeficiency virus type 1 preintegration complexes. *Proc Natl Acad Sci U S A* 89(14):6580–6584
 45. Medvedev SP, Shevchenko AI, Zakian SM (2010) Induced pluripotent stem cells: problems and advantages when applying them in regenerative medicine. *Acta Nat* 2(2):18–28
 46. Rao MS, Malik N (2012) Assessing iPSC reprogramming methods for their suitability in translational medicine. *J Cell Biochem* 113(10):3061–3068. <https://doi.org/10.1002/jcb.24183>
 47. Chang CW, Lai YS, Pawlik KM et al (2009) Polycistronic lentiviral vector for “hit and run” reprogramming of adult skin fibroblasts to induced pluripotent stem cells. *Stem Cells* 27(5):1042–1049. <https://doi.org/10.1002/stem.39>
 48. Soldner F, Hockemeyer D, Beard C et al (2009) Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. *Cell* 136(5):964–977. <https://doi.org/10.1016/j.cell.2009.02.013>
 49. Sommer CA, Stadtfeld M, Murphy GJ et al (2009) Induced pluripotent stem cell generation using a single lentiviral stem cell cassette. *Stem Cells* 27(3):543–549. <https://doi.org/10.1634/stemcells.2008-1075>
 50. Somers A, Jean JC, Sommer CA et al (2010) Generation of transgene-free lung disease-specific human induced pluripotent stem cells using a single excisable lentiviral stem cell cassette. *Stem Cells* 28(10):1728–1740. <https://doi.org/10.1002/stem.495>
 51. KUROYA M, ISHIDA N (1953) Newborn virus pneumonitis (type Sendai). II. The isolation of a new virus possessing hemagglutinin activity. *Yokohama Med Bull* 4(4):217–233
 52. Fusaki N, Ban H, Nishiyama A et al (2009) Efficient induction of transgene-free human pluripotent stem cells using a vector based on Sendai virus, an RNA virus that does not integrate into the host genome. *Proc Jpn Acad Ser B Phys Biol Sci* 85(8):348–362
 53. Ban H, Nishishita N, Fusaki N et al (2011) Efficient generation of transgene-free human induced pluripotent stem cells (iPSCs) by temperature-sensitive Sendai virus vectors. *Proc Natl Acad Sci U S A* 108(34):14234–14239. <https://doi.org/10.1073/pnas.1103509108>
 54. Schlaeger TM, Daheron L, Brickler TR et al (2015) A comparison of non-integrating reprogramming methods. *Nat Biotechnol* 33(1):58–63. <https://doi.org/10.1038/nbt.3070>

55. Sun TQ, Fenstermacher DA, Vos JM (1994) Human artificial episomal chromosomes for cloning large DNA fragments in human cells. *Nat Genet* 8(1):33–41. <https://doi.org/10.1038/ng0994-33>
56. Simpson K, McGuigan A, Huxley C (1996) Stable episomal maintenance of yeast artificial chromosomes in human cells. *Mol Cell Biol* 16(9):5117–5126
57. Westphal EM, Sierakowska H, Livanos E et al (1998) A system for shuttling 200-kb BAC/PAC clones into human cells: stable extra-chromosomal persistence and long-term ectopic gene activation. *Hum Gene Ther* 9(13):1863–1873. <https://doi.org/10.1089/hum.1998.9.13-1863>
58. Yu J, Hu K, Smuga-Otto K et al (2009) Human induced pluripotent stem cells free of vector and transgene sequences. *Science* 324(5928):797–801. <https://doi.org/10.1126/science.1172482>
59. Hu K, Yu J, Suknuntha K et al (2011) Efficient generation of transgene-free induced pluripotent stem cells from normal and neoplastic bone marrow and cord blood mononuclear cells. *Blood* 117(14):e109–e119. <https://doi.org/10.1182/blood-2010-07-298331>
60. Lin T, Ambasadhan R, Yuan X et al (2009) A chemical platform for improved induction of human iPSCs. *Nat Methods* 6(11):805–808. <https://doi.org/10.1038/nmeth.1393>
61. Yu J, Chau KF, Vodyanik MA et al (2011) Efficient feeder-free episomal reprogramming with small molecules. *PLoS One* 6(3):e17557. <https://doi.org/10.1371/journal.pone.0017557>
62. Chou BK, Mali P, Huang X et al (2011) Efficient human iPSC cell derivation by a non-integrating plasmid from blood cells with unique epigenetic and gene expression signatures. *Cell Res* 21(3):518–529. <https://doi.org/10.1038/cr.2011.12>
63. Okita K, Matsumura Y, Sato Y et al (2011) A more efficient method to generate integration-free human iPSC cells. *Nat Methods* 8(5):409–412. <https://doi.org/10.1038/nmeth.1591>
64. Warren L, Manos PD, Ahfeldt T et al (2010) Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell* 7(5):618–630. <https://doi.org/10.1016/j.stem.2010.08.012>
65. Goh PA, Caxaria S, Casper C et al (2013) A systematic evaluation of integration free reprogramming methods for deriving clinically relevant patient specific induced pluripotent stem (iPS) cells. *PLoS One* 8(11):e81622. <https://doi.org/10.1371/journal.pone.0081622>
66. Warren L, Ni Y, Wang J et al (2012) Feeder-free derivation of human induced pluripotent stem cells with messenger RNA. *Sci Rep* 2:657. <https://doi.org/10.1038/srep00657>
67. Merkle FT, Eggan K (2013) Modeling human disease with pluripotent stem cells: from genome association to function. *Cell Stem Cell* 12(6):656–668. <https://doi.org/10.1016/j.stem.2013.05.016>
68. Kajiwara M, Aoi T, Okita K et al (2012) Donor-dependent variations in hepatic differentiation from human-induced pluripotent stem cells. *Proc Natl Acad Sci U S A* 109(31):12538–12543. <https://doi.org/10.1073/pnas.1209979109>
69. Mills JA, Wang K, Paluru P et al (2013) Clonal genetic and hematopoietic heterogeneity among human-induced pluripotent stem cell lines. *Blood* 122(12):2047–2051. <https://doi.org/10.1182/blood-2013-02-484444>
70. Shao K, Koch C, Gupta MK et al (2013) Induced pluripotent mesenchymal stromal cell clones retain donor-derived differences in DNA methylation profiles. *Mol Ther* 21(1):240–250. <https://doi.org/10.1038/mt.2012.207>
71. Rouhani F, Kumasaka N, de Brito MC et al (2014) Genetic background drives transcriptional variation in human induced pluripotent stem cells. *PLoS Genet* 10(6):e1004432. <https://doi.org/10.1371/journal.pgen.1004432>
72. Kytala A, Moraghebi R, Valensisi C et al (2016) Genetic variability overrides the impact of parental cell type and determines iPSC differentiation potential. *Stem Cell Rep* 6(2):200–212. <https://doi.org/10.1016/j.stemcr.2015.12.009>
73. Bock C, Kiskinis E, Verstaappen G et al (2011) Reference maps of human ES and iPSC cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* 144(3):439–452. <https://doi.org/10.1016/j.cell.2010.12.032>
74. Boulting GL, Kiskinis E, Croft GF et al (2011) A functionally characterized test set of human induced pluripotent stem cells. *Nat Biotechnol* 29(3):279–286. <https://doi.org/10.1038/nbt.1783>
75. HD iPSC Consortium (2012) Induced pluripotent stem cells from patients with Huntington's disease show CAG-repeat-

- expansion-associated phenotypes. *Cell Stem Cell* 11(2):264–278. <https://doi.org/10.1016/j.stem.2012.04.027>
76. Cohen DE, Melton D (2011) Turning straw into gold: directing cell fate for regenerative medicine. *Nat Rev Genet* 12(4):243–252. <https://doi.org/10.1038/nrg2938>
77. Williams LA, Davis-Dusenbery BN, Eggen KC (2012) SnapShot: directed differentiation of pluripotent stem cells. *Cell* 149(5):1174–1174.e1. <https://doi.org/10.1016/j.cell.2012.05.015>
78. Lian X, Zhang J, Azarin SM et al (2013) Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating Wnt/beta-catenin signaling under fully defined conditions. *Nat Protoc* 8(1):162–175. <https://doi.org/10.1038/nprot.2012.150>
79. Yan Y, Shin S, Jha BS et al (2013) Efficient and rapid derivation of primitive neural stem cells and generation of brain subtype neurons from human pluripotent stem cells. *Stem Cells Transl Med* 2(11):862–870. <https://doi.org/10.5966/sctm.2013-0080>
80. Carlson C, Koonce C, Aoyama N et al (2013) Phenotypic screening with human iPS cell-derived cardiomyocytes: HTS-compatible assays for interrogating cardiac hypertrophy. *J Biomol Screen* 18(10):1203–1211. <https://doi.org/10.1177/1087057113500812>
81. Drawnel FM, Boccardo S, Prummer M et al (2014) Disease modeling and phenotypic drug screening for diabetic cardiomyopathy using human induced pluripotent stem cells. *Cell Rep* 9(3):810–821. <https://doi.org/10.1016/j.celrep.2014.09.055>
82. Slukvin II, Vodyanik MA, Thomson JA et al (2006) Directed differentiation of human embryonic stem cells into functional dendritic cells through the myeloid pathway. *J Immunol* 176(5):2924–2932
83. Erceg S, Lainez S, Ronaghi M et al (2008) Differentiation of human embryonic stem cells to regional specific neural precursors in chemically defined medium conditions. *PLoS One* 3(5):e2122. <https://doi.org/10.1371/journal.pone.0002122>
84. Prigodich AE, Seferos DS, Massich MD et al (2009) Nano-flares for mRNA regulation and detection. *ACS Nano* 3(8):2147–2152. <https://doi.org/10.1021/nm9003814>
85. Larsson HM, Lee ST, Rocco M et al (2012) Sorting live stem cells based on Sox2 mRNA expression. *PLoS One* 7(11):e49874. <https://doi.org/10.1371/journal.pone.0049874>
86. Tohyama S, Hattori F, Sano M et al (2013) Distinct metabolic flow enables large-scale purification of mouse and human pluripotent stem cell-derived cardiomyocytes. *Cell Stem Cell* 12(1):127–137. <https://doi.org/10.1016/j.stem.2012.09.013>
87. Nguyen HN, Byers B, Cord B et al (2011) LRRK2 mutant iPSC-derived DA neurons demonstrate increased susceptibility to oxidative stress. *Cell Stem Cell* 8(3):267–280. <https://doi.org/10.1016/j.stem.2011.01.013>
88. Israel MA, Yuan SH, Bardy C et al (2012) Probing sporadic and familial Alzheimer’s disease using induced pluripotent stem cells. *Nature* 482(7384):216–220. <https://doi.org/10.1038/nature10821>
89. Reinhardt P, Schmid B, Burbulla LF et al (2013) Genetic correction of a LRRK2 mutation in human iPSCs links parkinsonian neurodegeneration to ERK-dependent changes in gene expression. *Cell Stem Cell* 12(3):354–367. <https://doi.org/10.1016/j.stem.2013.01.008>
90. Ieda M, JD F, Delgado-Olguin P et al (2010) Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell* 142(3):375–386. <https://doi.org/10.1016/j.cell.2010.07.002>
91. Szabo E, Rampalli S, Risueno RM et al (2010) Direct conversion of human fibroblasts to multilineage blood progenitors. *Nature* 468(7323):521–526. <https://doi.org/10.1038/nature09591>
92. Vierbuchen T, Ostermeier A, Pang ZP et al (2010) Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* 463(7284):1035–1041. <https://doi.org/10.1038/nature08797>
93. Sekiya S, Suzuki A (2011) Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* 475(7356):390–393. <https://doi.org/10.1038/nature10263>
94. Ring KL, Tong LM, Balestra ME et al (2012) Direct reprogramming of mouse and human fibroblasts into multipotent neural stem cells with a single factor. *Cell Stem Cell* 11(1):100–109. <https://doi.org/10.1016/j.stem.2012.05.018>
95. Vierbuchen T, Wernig M (2011) Direct lineage conversions: unnatural but useful? *Nat Biotechnol* 29(10):892–907. <https://doi.org/10.1038/nbt.1946>
96. Yang C, Al-Aama J, Stojkovic M et al (2015) Concise review: cardiac disease modeling using induced pluripotent stem cells. *Stem*

- Cells 33(9):2643–2651. <https://doi.org/10.1002/stem.2070>
97. Nishi M, Akutsu H, Kudoh A et al (2014) Induced cancer stem-like cells as a model for biological screening and discovery of agents targeting phenotypic traits of cancer stem cell. *Oncotarget* 5(18):8665–8680
 98. Curry EL, Moad M, Robson CN et al (2015) Using induced pluripotent stem cells as a tool for modelling carcinogenesis. *World J Stem Cells* 7(2):461–469. <https://doi.org/10.4252/wjsc.v7.i2.461>
 99. Wiley LA, Burnight ER, Songstad AE et al (2015) Patient-specific induced pluripotent stem cells (iPSCs) for the study and treatment of retinal degenerative diseases. *Prog Retin Eye Res* 44:15–35. <https://doi.org/10.1016/j.preteyeres.2014.10.002>
 100. Zheng A, Li Y, Tsang SH (2015) Personalized therapeutic strategies for patients with retinitis pigmentosa. *Expert Opin Biol Ther* 15(3):391–402. <https://doi.org/10.1517/14712598.2015.1006192>
 101. Lysy PA, Weir GC, Bonner-Weir S (2012) Concise review: pancreas regeneration: recent advances and perspectives. *Stem Cells Transl Med* 1(2):150–159. <https://doi.org/10.5966/sctm.2011-0025>
 102. Abdelalim EM, Bonnefond A, Bennaceur-Griscelli A et al (2014) Pluripotent stem cells as a potential tool for disease modelling and cell therapy in diabetes. *Stem Cell Rev* 10(3):327–337. <https://doi.org/10.1007/s12015-014-9503-6>
 103. Peitz M, Jungverdorben J, Brustle O (2013) Disease-specific iPSC cell models in neuroscience. *Curr Mol Med* 13(5):832–841
 104. Crook JM, Wallace G, Tomaskovic-Crook E (2015) The potential of induced pluripotent stem cells in models of neurological disorders: implications on future therapy. *Expert Rev Neurother* 15(3):295–304. <https://doi.org/10.1586/14737175.2015.1013096>
 105. Goring HH, Curran JE, Johnson MP et al (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* 39(10):1208–1216. <https://doi.org/10.1038/ng2119>
 106. Winnier DA, Fourcaudot M, Norton L et al (2015) Transcriptomic identification of ADH1B as a novel candidate gene for obesity and insulin resistance in human adipose tissue in Mexican Americans from the Veterans Administration Genetic Epidemiology Study (VAGES). *PLoS One* 10(4):e0119941. <https://doi.org/10.1371/journal.pone.0119941>
 107. Pasca SP, Portmann T, Voineagu I et al (2011) Using iPSC-derived neurons to uncover cellular phenotypes associated with Timothy syndrome. *Nat Med* 17(12):1657–1662. <https://doi.org/10.1038/nm.2576>
 108. Prilutsky D, Palmer NP, Smedemark-Margulies N et al (2014) iPSC-derived neurons as a higher-throughput readout for autism: promises and pitfalls. *Trends Mol Med* 20(2):91–104. <https://doi.org/10.1016/j.molmed.2013.11.004>
 109. Wen Z, Nguyen HN, Guo Z et al (2014) Synaptic dysregulation in a human iPSC cell model of mental disorders. *Nature* 515(7527):414–418. <https://doi.org/10.1038/nature13716>
 110. Farra N, Zhang WB, Pasceri P et al (2012) Rett syndrome induced pluripotent stem cell-derived neurons reveal novel neurophysiological alterations. *Mol Psychiatry* 17(12):1261–1271. <https://doi.org/10.1038/mp.2011.180>
 111. Vaccarino FM, Urban AE, Stevens HE et al (2011) Annual research review: the promise of stem cell research for neuropsychiatric disorders. *J Child Psychol Psychiatry* 52(4):504–516. <https://doi.org/10.1111/j.1469-7610.2010.02348.x>
 112. Vaccarino FM, Stevens HE, Kocabas A et al (2011) Induced pluripotent stem cells: a new tool to confront the challenge of neuropsychiatric disorders. *Neuropharmacology* 60(7-8):1355–1363. <https://doi.org/10.1016/j.neuropharm.2011.02.021>
 113. Stevens HE, Mariani J, Coppola G et al (2012) Neurobiology meets genomic science: the promise of human-induced pluripotent stem cells. *Dev Psychopathol* 24(4):1443–1451. <https://doi.org/10.1017/S095457941200082X>
 114. Chae JI, Kim DW, Lee N et al (2012) Quantitative proteomic analysis of induced pluripotent stem cells derived from a human Huntington's disease patient. *Biochem J* 446(3):359–371. <https://doi.org/10.1042/BJ20111495>
 115. Szlachcic WJ, Switonski PM, Krzyzosiak WJ et al (2015) Huntington disease iPSCs show early molecular changes in intracellular signaling, the expression of oxidative stress proteins and the p53 pathway. *Dis Model Mech* 8(9):1047–1057. <https://doi.org/10.1242/dmm.019406>
 116. Weinberger DR (1987) Implications of normal brain development for the pathogenesis of schizophrenia. *Arch Gen Psychiatry* 44(7):660–669

117. White T, Anjum A, Schulz SC (2006) The schizophrenia prodrome. *Am J Psychiatry* 163(3):376–380
118. Gulsuner S, Walsh T, Watts AC et al (2013) Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* 154(3):518–529. <https://doi.org/10.1016/j.cell.2013.06.049>
119. Mariani J, Simonini MV, Palejev D et al (2012) Modeling human cortical development in vitro using induced pluripotent stem cells. *Proc Natl Acad Sci U S A* 109(31):12770–12775. <https://doi.org/10.1073/pnas.1202944109>
120. Nicholas CR, Chen J, Tang Y et al (2013) Functional maturation of hPSC-derived forebrain interneurons requires an extended timeline and mimics human neural development. *Cell Stem Cell* 12(5):573–586. <https://doi.org/10.1016/j.stem.2013.04.005>
121. Espuny-Camacho I, Michelsen KA, Gall D et al (2013) Pyramidal neurons derived from human pluripotent stem cells integrate efficiently into mouse brain circuits in vivo. *Neuron* 77(3):440–456. <https://doi.org/10.1016/j.neuron.2012.12.011>
122. Maroof AM, Keros S, Tyson JA et al (2013) Directed differentiation and functional maturation of cortical interneurons from human embryonic stem cells. *Cell Stem Cell* 12(5):559–572. <https://doi.org/10.1016/j.stem.2013.04.008>
123. Hu BY, Du ZW, Zhang SC (2009) Differentiation of human oligodendrocytes from pluripotent stem cells. *Nat Protoc* 4(11):1614–1622. <https://doi.org/10.1038/nprot.2009.186>
124. Wang S, Bates J, Li X et al (2013) Human iPSC-derived oligodendrocyte progenitor cells can myelinate and rescue a mouse model of congenital hypomyelination. *Cell Stem Cell* 12(2):252–264. <https://doi.org/10.1016/j.stem.2012.12.002>
125. Kim C, Wong J, Wen J et al (2013) Studying arrhythmogenic right ventricular dysplasia with patient-specific iPSCs. *Nature* 494(7435):105–110. <https://doi.org/10.1038/nature11799>
126. Maher B (2008) Personal genomes: the case of the missing heritability. *Nature* 456(7218):18–21. <https://doi.org/10.1038/456018a>
127. Manolio TA, Collins FS, Cox NJ et al (2009) Finding the missing heritability of complex diseases. *Nature* 461(7265):747–753. <https://doi.org/10.1038/nature08494>; [10.1038/nature08494](https://doi.org/10.1038/nature08494)
128. Franke A, McGovern DP, Barrett JC et al (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 42(12):1118–1125. <https://doi.org/10.1038/ng.717>
129. Cardno AG, Gottesman II (2000) Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. *Am J Med Genet* 97(1):12–17. [https://doi.org/10.1002/\(SICI\)1096-8628\(200021\)97:13.0.CO;2-U\[pii\]](https://doi.org/10.1002/(SICI)1096-8628(200021)97:13.0.CO;2-U[pii])
130. Sullivan PF, Kendler KS, Neale MC (2003) Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry* 60(12):1187–1192. <https://doi.org/10.1001/archpsyc.60.12.1187>
131. Visscher PM, Goddard ME, Derks EM et al (2012) Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. *Mol Psychiatry* 17(5):474–485. <https://doi.org/10.1038/mp.2011.65>
132. McGuire SE, McGuire AL (2008) Don't throw the baby out with the bathwater: enabling a bottom-up approach in genome-wide association studies. *Genome Res* 18(11):1683–1685. <https://doi.org/10.1101/gr.083584.108>
133. Tracy RP (2008) 'Deep phenotyping': characterizing populations in the era of genomics and systems biology. *Curr Opin Lipidol* 19(2):151–157. <https://doi.org/10.1097/MOL.0b013e3282f73893>
134. Vogelstein B, Lane D, Levine AJ (2000) Surfing the p53 network. *Nature* 408(6810):307–310. <https://doi.org/10.1038/35042675>
135. Chao EC, Lipkin SM (2006) Molecular models for the tissue specificity of DNA mismatch repair-deficient carcinogenesis. *Nucleic Acids Res* 34(3):840–852
136. Goh KI, Cusick ME, Valle D et al (2007) The human disease network. *Proc Natl Acad Sci U S A* 104(21):8685–8690
137. Lage K, Hansen NT, Karlberg EO et al (2008) A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc Natl Acad Sci U S A* 105(52):20870–20875. <https://doi.org/10.1073/pnas.0810772105>
138. Barshir R, Shwartz O, Smoly IY et al (2014) Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. *PLoS*

- Comput Biol 10(6):e1003632. <https://doi.org/10.1371/journal.pcbi.1003632>
139. Jenkinson CP, Goring HH, Arya R et al (2015) Transcriptomics in type 2 diabetes: bridging the gap between genotype and phenotype. *Genom Data* 8:25–36. <https://doi.org/10.1016/j.gdata.2015.12.001>
 140. Kim H, Kim JS (2014) A guide to genome engineering with programmable nucleases. *Nat Rev Genet* 15(5):321–334. <https://doi.org/10.1038/nrg3686>
 141. Boch J, Scholze H, Schornack S et al (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* 326(5959):1509–1512. <https://doi.org/10.1126/science.1178811>
 142. Wood AJ, Lo TW, Zeitler B et al (2011) Targeted genome editing across species using ZFNs and TALENs. *Science* 333(6040):307. <https://doi.org/10.1126/science.1207773>
 143. Sanjana NE, Cong L, Zhou Y et al (2012) A transcription activator-like effector toolbox for genome engineering. *Nat Protoc* 7(1):171–192. <https://doi.org/10.1038/nprot.2011.431>
 144. Cong L, Ran FA, Cox D et al (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* 339(6121):819–823. <https://doi.org/10.1126/science.1231143>
 145. Mali P, Yang L, Esvelt KM et al (2013) RNA-guided human genome engineering via Cas9. *Science* 339(6121):823–826. <https://doi.org/10.1126/science.1232033>
 146. Carroll D, Morton JJ, Beumer KJ et al (2006) Design, construction and in vitro testing of zinc finger nucleases. *Nat Protoc* 1(3):1329–1341
 147. Miller JC, Tan S, Qiao G et al (2011) A TALE nuclease architecture for efficient genome editing. *Nat Biotechnol* 29(2):143–148. <https://doi.org/10.1038/nbt.1755>
 148. Jinek M, Chylinski K, Fonfara I et al (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337(6096):816–821. <https://doi.org/10.1126/science.1225829>
 149. Musunuru K (2013) Genome editing of human pluripotent stem cells to generate human cellular disease models. *Dis Model Mech* 6(4):896–904. <https://doi.org/10.3824/stembook.1.94.1>
 150. Lombardo A, Genovese P, Beausejour CM et al (2007) Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery. *Nat Biotechnol* 25(11):1298–1306
 151. Suzuki K, Mitsui K, Aizawa E et al (2008) Highly efficient transient gene expression and gene targeting in primate embryonic stem cells with helper-dependent adenoviral vectors. *Proc Natl Acad Sci U S A* 105(37):13781–13786. <https://doi.org/10.1073/pnas.0806976105>
 152. Zou J, Maeder ML, Mali P et al (2009) Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells. *Cell Stem Cell* 5(1):97–110. <https://doi.org/10.1016/j.stem.2009.05.023>
 153. Hockemeyer D, Wang H, Kiani S et al (2011) Genetic engineering of human pluripotent cells using TALE nucleases. *Nat Biotechnol* 29(8):731–734. <https://doi.org/10.1038/nbt.1927>
 154. Li M, Suzuki K, Qu J et al (2011) Efficient correction of hemoglobinopathy-causing mutations by homologous recombination in integration-free patient iPSCs. *Cell Res* 21(12):1740–1744. <https://doi.org/10.1038/cr.2011.186>
 155. Sebastiano V, Maeder ML, Angstman JF et al (2011) In situ genetic correction of the sickle cell anemia mutation in human induced pluripotent stem cells using engineered zinc finger nucleases. *Stem Cells* 29(11):1717–1726. <https://doi.org/10.1002/stem.718>
 156. Soldner F, Laganieri J, Cheng AW et al (2011) Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. *Cell* 146(2):318–331. <https://doi.org/10.1016/j.cell.2011.06.019>
 157. Yusa K, Rashid ST, Strick-Marchand H et al (2011) Targeted gene correction of alpha1-antitrypsin deficiency in induced pluripotent stem cells. *Nature* 478(7369):391–394. <https://doi.org/10.1038/nature10424>
 158. Zou J, Mali P, Huang X et al (2011) Site-specific gene correction of a point mutation in human iPS cells derived from an adult patient with sickle cell disease. *Blood* 118(17):4599–4608. <https://doi.org/10.1182/blood-2011-02-335554>
 159. Lan F, Lee AS, Liang P et al (2013) Abnormal calcium handling properties underlie familial hypertrophic cardiomyopathy pathology in patient-specific induced pluripotent stem cells. *Cell Stem Cell* 12(1):101–113. <https://doi.org/10.1016/j.stem.2012.10.010>

Chapter 3

Development of Targeted Therapies Based on Gene Modification

Taylor M. Benson, Fatjon Leti, and Johanna K. DiStefano

Abstract

With the advent of next-generation sequencing (NGS) and the demand for a personalized healthcare system, the fields of precision medicine and gene therapy are advancing in new directions. There is a push to identify genes that contribute to disease development, either alone or in conjunction with other genes or environmental factors, and then design targeted therapies based on this knowledge, rather than the traditional approach of treating generalized symptoms with pharmaceuticals in a one-size-fits-all manner. Identification of genes that contribute to disease pathogenesis and progression is critical for the maturation of the precision medicine field. Concomitant with a better understanding of disease pathology, precision medicine approaches can be adopted with greater confidence and are expected to lead to a new standard for clinical practice. In this chapter, we provide a brief introduction to precision medicine, discuss the importance of identifying genes and genetic variants that contribute to disease development and progression, offer examples of approaches that can be applied to treat specific diseases, and present some of the current challenges and limitations of precision medicine.

Key words Precision medicine, Personalized medicine, Gene therapy, Pharmacogenomics, NGS, GWAS

1 Introduction

Healthcare in the USA has traditionally applied a “one-size-fits-all” approach to the treatment of disease, regardless of its etiology. Due to heterogeneity in disease presentation, as well as overlapping manifestations among different disorders, precise definition of phenotypic abnormalities in patients can be challenging. Many patients fall through cracks in the healthcare system due to errors in diagnosis, lack of accurate prognostic tools, or failure to account for differential responses to pharmaceutical treatments. Improved strategies to take patient-specific factors into consideration underlie the practice of precision medicine. This recently constructed model for clinical practice seeks to treat and prevent disease by addressing

variability in genetic, environmental, and lifestyle factors in the individual patient.

In theory, precision medicine tailors medical treatment for individual patients based upon shared characteristics among a group of individuals. For example, when patients are first diagnosed with type 2 diabetes or hypercholesterolemia, they are frequently prescribed metformin or statins, respectively, regardless of the etiology of the condition. While metformin and statins are effective for a large proportion of patients, these pharmaceuticals do not yield the desired effects for many individuals and cause unpleasant side effects in others. Accounting for individual factors contributing to disease pathogenesis and pharmacogenetic response would allow the most effective pharmacological strategy to be implemented upon diagnosis, thereby restoring the patient to health more quickly and effectively. Likewise, strategies to monitor disease risk and progression to more clinically severe manifestations of the disease using biomarkers remain largely undeveloped, despite the potential for disease prevention or reversal [1].

With the advancement of technologies such as next generation sequencing (NGS) techniques, a shift in our understanding of the role of genetic variants in the development of human diseases has been experienced. While these platforms have taken several years to yield robust results, they have produced a significant amount of data with respect to long-term health and disease. The emerging molecular biology methods and technologies discussed in this book, in conjunction with pharmacogenomics studies, will be crucial to more effectively diagnose and treat patients [2]. Because of the impact that the practice of precision medicine is expected to have on the treatment and prevention of disease, we have focused on integrating knowledge obtained from genetic studies and approaches to correct gene deficiencies and dysfunction in this chapter. We also discuss some of the challenges and limitations of precision medicine as a sustainable paradigm for healthcare.

2 The Importance of Taking into Account Genetic and Environmental Factors in the Prevention and Treatment of Human Disease

Theoretically, the practice of precision medicine takes into account genetic, environmental, and lifestyle factors of individual patients. The underlying premise of this clinical approach suggests that the better characterized genes and genetic variants are in the pathogenesis and progression of disease, the more effectively treatment strategies can be administered to patients. In contrast to rare monogenic diseases, which arise due to a defect in a single gene, most common disorders, such as heart disease, neurocognitive problems, and diabetes, result from the combination of many different factors, including genetic predisposition. For example, levels

of circulating thyroid hormone and thyroid stimulating hormone (TSH) are strongly dependent on genetic factors [3]; the development of hypothyroidism and hyperthyroidism is mediated by genetic variants, including those in the genes encoding phosphodiesterase 8B, iodothyronine deiodinase 1, F-actin-capping protein subunit beta, and the TSH receptor [3]. Genetic variation also underlies differential response to pharmacological treatments for hypothyroidism. Approximately, 10–15% of patients who take levothyroxine (synthetic T4 hormone) do not benefit from this drug because they cannot effectively convert T4 to T3 due to genetic, nutritional, or hormonal factors [4]. Knowledge of genes and genetic variants that contribute to disease pathogenesis is expected to facilitate the development of more specific strategies to administer the most effective course of treatment with minimal side effects.

Gene–environment interactions (GxE) address how different genotypes respond to the same environmental stimuli in dissimilar ways. In the practice of precision medicine, GxE interactions may provide insights into the manner in which the effect of genetic variation might be attenuated through lifestyle interventions to prevent, treat, or manage the disease. In the case of thyroid disorders, for example, 33% of thyroid hormone and TSH levels are attributable to GxE [3]. Thus, although a patient may have a genetic predisposition to altered thyroid function, implementation of specific lifestyle changes may mitigate its effects. For example, poor nutrition and sedentary behavior are known to play a role in the progression of some thyroid diseases, while plant-based diets are associated with lower risk of developing hypothyroidism [5].

Despite the importance of GxE interactions in the development of disease, quantifying the manner in which environmental exposures modulates genetic risk remains a difficult task. To address the impact of environmental covariates in an *in vitro* system, a recent study focused on changes in the cellular environment in response to naturally occurring environmental exposures [6]. Using allele-specific expression to characterize genetic effects on the transcriptional response to 50 treatments in five cell types, nearly 1500 genes with allele-specific expression were identified and 215 genes were found to participate in GxE interactions. Genes responding to environmental changes were more likely to be identified in genome-wide association studies (GWAS) and 49% of these genes were associated with complex traits. [6]. Examination of per variant heritability for 18 complex traits using a mixed model approach revealed that regulation of HDL-cholesterol, total cholesterol, and mean corpuscular hemoglobin levels was largely controlled by GxE interactions.

In this study, genes showing allele-specific expression induced by environmental perturbations were also assessed and 75 variants spanning 60 genes were identified, 28 of which were associated with a phenotype in the GWAS catalog [6]. Results of this analysis

suggested that caffeine upregulates expression of gastric inhibitory polypeptide receptor (*GIPR*), which has been previously linked with obesity and related traits. Higher gene expression and allele-specific expression in response to caffeine was identified for the major allele of a *GIPR* variant, rs5380, and this allele is located on a haplotype associated with normal body mass index [7], suggesting that caffeine may protect against the development of obesity through its effect on gene expression and allele-specific expression in *GIPR*. Similarly, in an analysis of condition-specific changes in allele-specific expression, four genes were associated with complex traits, including SAMM50 sorting and assembly machinery component (*SAMM50*), which responded to copper, endoplasmic reticulum aminopeptidase 1 (*ERAP1*), which responded to selenium, golgi SNAP receptor complex member 2 (*GOSR2*), which responded to mono-n-butyl-phthalate, and lysosomal associated membrane protein 3 (*LAMP3*), which also responded to selenium. A variant in *LAMP3*, rs16833703, preferentially expressed the alternative allele, which is located on a protective haplotype for Parkinson's disease [8]. In combination with earlier studies showing reduced selenium levels in individuals with Parkinson's disease, these results suggest selenium may be a beneficial therapeutic adjunct through its effects on allelic expression of *LAMP3*.

Epigenetic components, such as DNA methylation, histone modification, and regulatory noncoding RNAs (ncRNAs), can also be altered by environmental factors. For example, the activity of *CYP1A2*, which plays a key role in both drug metabolism and synthesis of cholesterol and lipids, is regulated by smoking and diet through mechanisms involving promoter hypermethylation [9]. Likewise, diet and exercise have been shown to alter microRNA (miRNA) profiles, which may lead to changes in the development and progression of diseases such as diabetes, cardiovascular disease, and obesity through regulation of adipogenesis and lipid metabolism [10, 11].

3 Methods for Correcting Gene Deficiencies and Dysfunction

Gene therapy treats dysfunctional genes that underlie human disease using targeted delivery of nucleic acids. The approach is applied when levels of protein products of affected genes are produced in inadequate amounts, by restoring levels of a deficient gene product or inhibiting an overexpressed or defective transcript. In essence, gene therapy can be defined as a set of approaches that utilizes endogenous transcription or translation through the transfer of exogenous genetic material. Gene therapy can be applied to somatic or germ cells, however in the USA, the use of gene therapy is limited to somatic cells to avoid passing side effects to future generations.

3.1 Gene Therapy Approaches

Gene therapies can include gene augmentation, gene inhibition, genome editing, and destruction of specific cells. Gene augmentation adds functional copies of a deficient gene, either transiently through the addition of DNA plasmid or mRNA transcript, or sustainably, through insertion of a transgene into the patient's genome. Lentiviruses and oncoretroviruses are convenient vectors for transgene insertion because they provide efficient cell entry and integration. Alternatively, transposons can be used for transgene insertion. Transposons are excellent choices for large-scale production and biosafety, but do not show the same efficiency in cell entry as virus-mediated approaches. Additionally, programmable endonucleases, discussed below, can direct the transposon to insert the transgene into specific regions in the genome [12, 13].

For inhibition of overexpressed or defective disease-causing genes, the mutated sequence is replaced with the wild type version through editing by CRISPR/Cas9 [14] or downregulated by the fine-tuning of antisense oligonucleotides (ASOs) [15], locked nucleic acids (LNAs) [16] or short hairpin RNAs (shRNAs), small interfering RNAs (siRNAs), miRNAs, and antagomirs [17, 18].

For diseases like cancer, the most effective therapeutic approach is to target and eliminate entire populations of cells. Such mass elimination can be achieved using cell-specific insertion of a gene that produces a cytotoxic product [19]. Alternatively, transgenes expressing a membrane protein that labels the cell to be attacked by the immune system can be used to eliminate specific cells [20]. For example, T cells are modified to express chimeric antigen receptors to elicit a T cell response to the cancer cells expressing a specific antigen [21].

Traditional gene therapy methods add exogenous genetic material to the nucleus or genome at a nonspecific location, which may result in unintended side effects. As mentioned above, programmable endonucleases, such as zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and clustered regularly interspaced short palindromic repeats/CRISPR-associated nuclease 9 (CRISPR/Cas9) accurately bind specific DNA sequences. With the incorporation of nucleases, transposases, and other enzymes, these endonucleases are able to elicit many different precise gene-editing effects. CRISPR/Cas9 is derived from microbial adaptive immune defense system and can be applied in mammalian cells as a genome-editing tool [22]. The CRISPR/Cas9 system requires a nuclease Cas9, a single guide RNA (sgRNA) derived from CRISPR RNA (crRNA), and transacting CRISPR RNA. If there is a protospacer-adjacent motif (PAM) on the DNA, the Cas9 is guided to the complementary sequence of the sgRNA by base pairing. The Cas9 generates double-strand breaks (DSBs) at the desired sites. If homologous sequences are available, breaks are repaired by homologous directed repair (HDR) leading to precise gene replacement or correction; if not,

they are repaired by nonhomologous end-joining (NHEJ), potentially inducing small insertion or deletion (indel) mutations. Although CRISPR is limited by the presence of a PAM sequence—a two to six base pair sequence on the nontarget strand necessary for sgRNA recognition—this is an easier approach to use compared to other programmable endonucleases as it is smaller than ZFN and TALEN guiding proteins and needs only a single complementary sgRNA [23].

Because genome editing induces permanent changes to the genome, the target specificity of this technology needs to be fully addressed prior to use in clinical trials. Off-target Cas9 activity can disrupt expression of important genes and even lead to the development of cancer. In China, CRISPR/Cas9 was tested for editing of the hemoglobin subunit beta (*HBB*) gene as a potential treatment for beta-thalassemia using human tripronuclear zygotes [24]. While CRISPR/Cas9 was found to effectively cleave *HBB*—28 of the 54 embryos showed cleavage—the efficiency of the homologous recombination directed repair was low, the edited embryos were mosaics, and some zygotes showed off-target cleavage. The endogenous hemoglobin subunit delta (*HBD*) gene, which is homologous with *HBB*, competed with *HBB* as a repair template leading to point mutations in seven zygotes. Despite an eight base pair mismatch, CRISPR/Cas9 was also found to target the unrelated complement C1q C chain (*C1QC*) locus. These results reveal that issues related to fidelity and specificity must be addressed before CRISPR/Cas9 gene editing can be safely used in clinical applications.

3.2 Delivery of Nucleic Acids in Gene Therapy

One of the main obstacles to widespread clinical implementation of gene therapy is delivery of transgenes. Nucleic acids are large and negatively charged so a vector or “vehicle” for transporting them into the cell is necessary. Once in the cell, the genetic information enters the nucleus either as an extrachromosomal episome or through genome integration, and must be activated to have an effect. The delivery vehicle is either applied *in vivo* or *ex vivo* to modify cells, depending upon the disease. *Ex vivo* gene therapy is limited to dividing cells, is less immunogenic, and transgene expression can be measured before implanting the cells back to the patient. *In vivo* therapy is faster and easier, but there is a greater chance of immune response and transfection of off-target cells. A summary of current nucleic acid delivery methods is shown in Table 1.

For *in vivo* gene therapy, viruses are the most commonly used vectors. Viruses have evolved to enter mammalian cells and deliver genetic material to the nucleus, and therefore have high transfection efficiency. However, viruses have limited DNA packaging capacity [25] and can cause mutagenesis, carcinogenesis [26], undesirable immune response [27], and nonspecific insertion.

Table 1
Methods for gene delivery

Delivery method	Advantages	Disadvantages	Selected clinical trials	Ref
Viral vectors	High transfection efficiency	Limited DNA packaging capacity; Insertional mutation; Immune response; Non-specific insertion	Phase I/II; Completed December 2015; Pompe Disease treated with adeno-associated viral vector containing acid alpha-glucosidase (GAA). NCT009763532.	[25–27, 57]
Liposomal vectors	Low risk of mutagenesis; Delivery of RNA and DNA	Quickly degraded; Challenging entry to nucleus	Phase I; Completed April 2011; Non-Small-Cell Lung Cancer treated with DOTAP:Chol-fus1 liposome complex for delivery of fus1. NCT00059605.	[30]
Nanobombs	Low risk of mutagenesis; Delivery of RNA and DNA	Tissues must be accessible to be subjected to NIR	N/A	[31]
Ex vivo transplant	Low immune response; Low risk of blood proteins dismantling vector; Transgene expression is easily measured; Low risk of off-target cells affected	Limited to dividing cells; Labor intensive	Phase I; Completed July 2011; XSCID treated with onco-retroviral vector treated stem cells with gene for delivery of CD34 + gene. NCT00028236. Phase I; Completed November 2014; Osteoarthritis treated with ex vivo cultured adult allogenic mesenchymal stem cells. NCT01586312.	[32]

Non-viral, or synthetic, vectors are less likely to cause immune response or mutagenesis, have the potential to deliver more genetic material, and are safer to synthesize [28]. However, nonviral delivery vectors are typically plasmids, which are less efficient and more quickly degraded.

The most common nonviral vectors are liposomal vectors, which are spherical vesicles of cationic phospholipids that bind and transport nucleic acids into the cell [29]. The lipid bilayer of the liposome facilitates transport through the cell membrane. As a result of this process, the liposome enters the cell as an endosome;

however, once inside, the endosome must be dismantled to release nucleic acids to the cytosol. Typically, the lipids of the liposome trigger endosome destruction in response to changes in pH [30]. Nanobombs, which are nanoparticles containing indocyanine green and ammonium bicarbonate, have been developed to specifically deliver miRNA to cells [31]. Nanobombs can be used to destroy tumors *in vivo* with minimal side effects.

Ex vivo cell transplants can be used for diseases that primarily affect a specific cell type. In this approach, affected cells are removed from the patient, cultured, and then modified *in vitro* with a viral or nonviral vector. Successfully modified cells are then delivered back to the patient. This approach eliminates the need for gene delivery vectors, which may lead to improvements in safety, cell specificity, and efficacy [32].

Stem cells are pluripotent and self-renewing, and are therefore the best cells for *ex vivo* gene therapy. Hematopoietic stem cells (HSCs) are precursors to T cells, which are involved in immune response. T cell receptors (TCRs) bind specific antigens on target cells to mediate their destruction, and have been successfully manipulated to express receptors allowing recognition of target surface molecules on tumors [33]. Regulatory T cells are responsible for turning off the body's immune response and have also been modified *ex vivo* to treat autoimmune diseases [34]. Clinically, HSCs have been used to treat adrenoleukodystrophy [35] and adenosine deaminase-deficient severe combined immune deficiency [36].

Induced pluripotent stem cells (iPSCs) are a type of stem cell that can be derived from somatic cells by introducing a cocktail of transcription factors [37], small molecule compounds [38], or alternate vectors [39–41] to somatic cells [42]. Because iPSCs are derived from somatic tissue, individuals can contribute to the development of their own pluripotent cell lines. Ocular cells, including corneal epithelial-like cells, retinal pigment epithelium photoreceptors, and retinal ganglion cells have all been derived from iPSCs, and there are at least seven clinical trials of iPSC or embryonic stem cell studies in ocular diseases currently ongoing [43]. The first clinical study where iPSCs were used was for the treatment of macular degeneration via retinal pigment epithelial iPSC derived cells. In that study, skin cells were taken from a patient with retinal damage from age-related macular degeneration, reprogrammed into iPSCs, differentiated into retinal tissue, and then transplanted them into the eye [44]. While the procedure did not improve the patient's vision, it did prove to be safe and it halted further progression of the disease [45].

Patient-derived iPSCs can also be used in gene therapy approaches. For example, application of exon knock-in in iPSCs derived from Duchenne muscular dystrophy patients, followed by differentiation of corrected iPSCs to skeletal muscle cells,

successfully produced full-length dystrophin, the dysregulated gene that causes the disease [46]. Similarly, correction of beta-thalassemia mutations in patient-derived iPSCs promoted hematopoietic differentiation in mice [47].

4 Pharmacogenomics

The field of pharmacogenomics embodies the study of the role of the human genome in determining response to pharmacological therapy. Individual response to drugs can be mitigated by genetic makeup, particularly by genetic variation in drug-metabolizing enzymes and transporter proteins [48]. Genetic variants can predict pharmacological response in terms of drug absorption, distribution, metabolism, or elimination. With the advent of broad-range sequencing technologies, the field of pharmacogenomics has rapidly progressed. Patients are typically grouped by phenotype for a certain drug reaction or drug efficacy, DNA is analyzed, and candidate SNPs are predicted.

A wide array of pharmaceuticals such as antidepressants, cholesterol and lipid-lowering molecules, and cancer treatments have been investigated in pharmacogenomics studies. For example, the metabolism of warfarin, a commonly used blood thinner, is mediated by variants in the *CYP2C9* and *VKORC1* genes, and patients with these variants require a lower dosage of the drug or treatment with an alternative anticoagulant [49]. Likewise, variants in the *CYP2D6* gene increase codeine sensitivity, and affected individuals should be instead treated with morphine or nonopioid analgesics [49]. Although genetic variants underlie differential drug response, factors such as age, sex, and environmental exposures may also contribute to all phases of drug metabolism and should be taken into consideration when assessing pharmacological outcomes of disease management [50].

5 Challenges and Limitations of Precision Medicine as a Sustainable Paradigm for Healthcare

Precision medicine has been important for determining the efficacy of disease treatments. For example, the erb-b2 receptor tyrosine kinase 2 (HER2) is overexpressed in approximately 20% of breast cancers. For HER2-positive breast cancer, treatment with the HER2-targeted antibody, trastuzumab (i.e., Herceptin) is effective. However, for patients without elevated HER2 levels, this therapy yields little benefit [51]. In cystic fibrosis, traditional therapies have focused on the secondary consequences of the disease without taking genetic etiology into consideration. Cystic fibrosis is caused

by mutations in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene, each of which exerts different effects on the protein product. For example, in some cases, *CFTR* has impaired activity but retains normal targeting to the cell surface, while in others, the protein is not properly incorporated in the cell membrane. Thus, there is substantial heterogeneity in the etiology of the disease. Ivacaftor is a key pharmacological agent used to treat cystic fibrosis, but because it acts by increasing the opening time of the *CFTR* channel, the drug is only effective in those patients in whom the protein has reached the cell surface. For patients who have no *CFTR* incorporation at the cell surface, ivacaftor can be combined with another drug, lumacaftor, to improve delivery of the channel to the cell surface [49]. Increasing clinical awareness of genetic variability in drug response is leading to better quality of life for cystic fibrosis patients.

Progression in the field of precision medicine depends largely on the generation and analysis of “big data,” which refers to extremely large data sets that can be computationally analyzed to identify patterns and associations. The healthcare industry is undergoing a paradigm shift from the reactive model of treating symptoms to a more predictive model based on such data [52]. A major challenge of precision medicine is incorporating all layers of disease. Currently, single prognostic gene biomarkers such as *HER2* and *CFTR* are used to diagnose and efficiently treat breast cancer and cystic fibrosis, as mentioned above. However, interactions between biomarkers and environmental exposures, which may be more of the norm rather than the exception, are more difficult to evaluate [53]. While the field of precision medicine aims to integrate data from “omics” platforms (e.g., genomic, transcriptomic, proteomic, and metabolomic), environmental factors, and patient information, doing so presents a notable challenge [54]. Similarly, big data analysis may reveal that unrelated conditions share common dysregulated targets, yet converting different types of data into a simple output to determine what treatment will bring dysregulated pathways to a healthy condition remains a difficult endeavor [55].

6 Conclusions

Ever since the announcement of Precision Medicine Initiative by the US President Barack Obama in 2015, efforts to identify genetic and pathophysiological mechanisms of many human diseases using high throughput technologies has escalated. For precision medicine to deliver results, enhancements in genomics technologies and analytical strategies are needed, both of which have taken an upward shift due to the recent scientific, technological, and social developments. However, treating genetically based diseases via gene therapy requires years of investigation and the advancement

of delivery methods before it can be applied to clinical practice. For instance, issues related to transgene delivery must be addressed before treatments such as gene editing and ex vivo transplant therapy can be realized. On the other hand, decreasing costs associated with NGS technologies have quickly resulted in the output of a tremendous amount of “big data,” which has consequently led to the development of many new data analysis methodologies and computing capacities [56]. Cutting edge molecular biology methods continue to emerge, allowing more rapid and sophisticated means to for functional validation of genetic variants, as well as dysregulated coding and noncoding RNA transcripts. Combined, these approaches are expected to eventually provide a foundation for enhanced precision in the diagnosis and clinical management of human disease.

References

1. Szeffler SJ, Martin RJ (2010) Lessons learned from variation in response to therapy in clinical trials. *J Allergy Clin Immunol* 125:285–292. quiz 293–284
2. Yu H, Zhang VW (2015) Precision medicine for continuing phenotype expansion of human genetic diseases. *Biomed Res Int* 2015:745043
3. Panicker V (2011) Genetics of thyroid function and disease. *Clin Biochem Rev* 32:165–175
4. McAninch EA, Bianco AC (2016) The history and future of treatment of hypothyroidism. *Ann Intern Med* 164:50–56
5. Tonstad S, Nathan E, Oda K, Fraser G (2013) Vegan diets and hypothyroidism. *Nutrients* 5:4642–4652
6. Moyerbrailean GA, Richards AL, Kurtz D, Kalita CA, Davis GO, Harvey CT, Alazizi A, Wata D, Sorokin Y, Hauff N, Zhou X, Wen X, Pique-Regi R, Luca F (2016) High-throughput allele-specific expression across 250 environmental conditions. *Genome Res* 26:1627–1638
7. Wen W, Cho YS, Zheng W, Dorajoo R, Kato N, Qi L, Chen CH, Delahanty RJ, Okada Y, Tabara Y, Gu D, Zhu D, Haiman CA, Mo Z, Gao YT, Saw SM, Go MJ, Takeuchi F, Chang LC, Kokubo Y, Liang J, Hao M, Le Marchand L, Zhang Y, Hu Y, Wong TY, Long J, Han BG, Kubo M, Yamamoto K, Su MH, Miki T, Henderson BE, Song H, Tan A, He J, Ng DP, Cai Q, Tsunoda T, Tsai FJ, Iwai N, Chen GK, Shi J, Xu J, Sim X, Xiang YB, Maeda S, Ong RT, Li C, Nakamura Y, Aung T, Kamatani N, Liu JJ, Lu W, Yokota M, Seielstad M, Fann CS, A. T. C. Genetic Investigation of, J. Y. Wu, J. Y. Lee, F. B. Hu, T. Tanaka, E. S. Tai, and X. O. Shu (2012) Meta-analysis identifies common variants associated with body mass index in east Asians. *Nat Genet* 44:307–311
8. Do CB, Tung JY, Dorfman E, Kiefer AK, Drabant EM, Francke U, Mountain JL, Goldman SM, Tanner CM, Langston JW, Wojcicki A, Eriksson N (2011) Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson’s disease. *PLoS Genet* 7:e1002141
9. Miyajima A, Furihata T, Chiba K (2009) Functional analysis of GC Box and its CpG methylation in the regulation of CYP1A2 gene expression. *Drug Metab Pharmacokinet* 24:269–276
10. Flowers E, Won GY, Fukuoka Y (2015) MicroRNAs associated with exercise and diet: a systematic review. *Physiol Genomics* 47:1–11
11. Peng Y, Yu S, Li H, Xiang H, Peng J, Jiang S (2014) MicroRNAs: emerging roles in adipogenesis and obesity. *Cell Signal* 26:1888–1896
12. Owens JB, Mauro D, Stoytchev I, Bhakta MS, Kim MS, Segal DJ, Moisyadi S (2013) Transcription activator like effector (TALE)-directed piggyBac transposition in human cells. *Nucleic Acids Res* 41:9197–9207
13. Voigt K, Gogol-Doring A, Miskey C, Chen W, Cathomen T, Izsvak Z, Ivics Z (2012) Retargeting sleeping beauty transposon insertions by engineered zinc finger DNA-binding domains. *Mol Ther* 20:1852–1862
14. Xue HY, Zhang X, Wang Y, Xiaojie L, Dai WJ, Xu Y (2016) In vivo gene therapy potentials of CRISPR-Cas9. *Gene Ther* 23:557–559
15. Magner D, Biala E, Lisowiec-Wachnicka J, Kierzek E, Kierzek R (2015) A tandem oligonucleotide approach for SNP-selective RNA

- degradation using modified antisense oligonucleotides. *PLoS One* 10:e0142139
16. Grunweller A, Hartmann RK (2007) Locked nucleic acid oligonucleotides: the next generation of antisense agents? *BioDrugs* 21:235–243
 17. Kobayashi H, Tomari Y (2016) RISC assembly: coordination between small RNAs and Argonaute proteins. *Biochim Biophys Acta* 1859:71–81
 18. Krutzfeldt J, Rajewsky N, Braich R, Rajeev KG, Tuschl T, Manoharan M, Stoffel M (2005) Silencing of microRNAs in vivo with ‘antagomirs’. *Nature* 438:685–689
 19. Kolypetri P, King J, Larijani M, Carayanniotis G (2015) Genes and environment as predisposing factors in autoimmunity: acceleration of spontaneous thyroiditis by dietary iodide in NOD.H2(h4) mice. *Int Rev Immunol* 34:542–556
 20. Zhang E, Xu H (2017) A new insight in chimeric antigen receptor-engineered T cells for cancer immunotherapy. *J Hematol Oncol* 10:1
 21. Posey AD Jr, Schwab RD, Boesteanu AC, Steentoft C, Mandel U, Engels B, Stone JD, Madsen TD, Schreiber K, Haines KM, Cogdill AP, Chen TJ, Song D, Scholler J, Kranz DM, Feldman MD, Young R, Keith B, Schreiber H, Clausen H, Johnson LA, June CH (2016) Engineered CAR T cells targeting the cancer-associated tn-glycoform of the membrane mucin MUC1 control adenocarcinoma. *Immunity* 44:1444–1454
 22. 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
 23. Doudna JA, Charpentier E (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* 346:1258096
 24. Liang P, Xu Y, Zhang X, Ding C, Huang R, Zhang Z, Lv J, Xie X, Chen Y, Li Y, Sun Y, Bai Y, Songyang Z, Ma W, Zhou C, Huang J (2015) CRISPR/Cas9-mediated gene editing in human trippronuclear zygotes. *Protein Cell* 6:363–372
 25. Thomas CE, Ehrhardt A, Kay MA (2003) Progress and problems with the use of viral vectors for gene therapy. *Nat Rev Genet* 4:346–358
 26. Baum C, Kustikova O, Modlich U, Li Z, Fehse B (2006) Mutagenesis and oncogenesis by chromosomal insertion of gene transfer vectors. *Hum Gene Ther* 17:253–263
 27. Bessis N, GarciaCozar FJ, Boissier MC (2004) Immune responses to gene therapy vectors: influence on vector function and effector mechanisms. *Gene Ther* 11(Suppl 1):S10–S17
 28. Pack DW, Hoffman AS, Pun S, Stayton PS (2005) Design and development of polymers for gene delivery. *Nat Rev Drug Discov* 4:581–593
 29. Tseng WC, Haselton FR, Giorgio TD (1997) Transfection by cationic liposomes using simultaneous single cell measurements of plasmid delivery and transgene expression. *J Biol Chem* 272:25641–25647
 30. Farhood H, Serbina N, Huang L (1995) The role of dioleoyl phosphatidylethanolamine in cationic liposome mediated gene transfer. *Biochim Biophys Acta* 1235:289–295
 31. Wang H, Agarwal P, Zhao S, Yu J, Lu X, He X (2016) A near-infrared laser-activated “Nanobomb” for breaking the barriers to microRNA delivery. *Adv Mater* 28:347–355
 32. Naldini L (2011) Ex vivo gene transfer and correction for cell-based therapies. *Nat Rev Genet* 12:301–315
 33. Gschwend E, De Oliveira S, Kohn DB (2014) Hematopoietic stem cells for cancer immunotherapy. *Immunol Rev* 257:237–249
 34. Jethwa H, Adami AA, Maher J (2014) Use of gene-modified regulatory T-cells to control autoimmune and alloimmune pathology: is now the right time? *Clin Immunol* 150:51–63
 35. Cartier N, Hacein-Bey-Abina S, Bartholomae CC, Veres G, Schmidt M, Kutschera I, Vidaud M, Abel U, Dal-Cortivo L, Caccavelli L, Mahlaoui N, Kiermer V, Mittelstaedt D, Bellesme C, Lahlou N, Lefrere F, Blanche S, Audit M, Payen E, Leboulch P, l’Homme B, Bougneres P, Von Kalle C, Fischer A, Cavazzana-Calvo M, Aubourg P (2009) Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science* 326:818–823
 36. Candotti F, Shaw KL, Muul L, Carbonaro D, Sokolic R, Choi C, Schurman SH, Garabedian E, Kesserwan C, Jagadeesh GJ, Fu PY, Gschwend E, Cooper A, Tisdale JF, Weinberg KI, Crooks GM, Kapoor N, Shah A, Abdel-Azim H, Yu XJ, Smogorzewska M, Wayne AS, Rosenblatt HM, Davis CM, Hanson C, Rishi RG, Wang X, Gjertson D, Yang OO, Balamurugan A, Bauer G, Ireland JA, Engel BC, Podsakoff GM, Hershfield MS, Blaese RM, Parkman R, Kohn DB (2012) Gene therapy for adenosine deaminase-deficient severe combined immune deficiency: clinical comparison of retroviral vectors and treatment plans. *Blood* 120:3635–3646

37. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126:663–676
38. Shi Y, Despons C, Do JT, Hahm HS, Scholer HR, Ding S (2008) Induction of pluripotent stem cells from mouse embryonic fibroblasts by Oct4 and Klf4 with small-molecule compounds. *Cell Stem Cell* 3:568–574
39. Maherali N, Ahfeldt T, Rigamonti A, Utikal J, Cowan C, Hochedlinger K (2008) A high-efficiency system for the generation and study of human induced pluripotent stem cells. *Cell Stem Cell* 3:340–345
40. Okita K, Nakagawa M, Hyenjong H, Ichisaka T, Yamanaka S (2008) Generation of mouse induced pluripotent stem cells without viral vectors. *Science* 322:949–953
41. Zhou W, Freed CR (2009) Adenoviral gene delivery can reprogram human fibroblasts to induced pluripotent stem cells. *Stem Cells* 27:2667–2674
42. Takahashi K, Yamanaka S (2013) Induced pluripotent stem cells in medicine and biology. *Development* 140:2457–2461
43. Wu N, Doorenbos M, Chen DF (2016) Induced pluripotent stem cells: development in the ophthalmologic field. *Stem Cells Int* 2016:2361763
44. Reardon S, Cyranoski D (2014) Japan stem-cell trial stirs envy. *Nature* 513:287–288
45. Normile D (2017) iPSC cell therapy reported safe. *Science* 355:1109–1110
46. Li HL, Fujimoto N, Sasakawa N, Shirai S, Ohkame T, Sakuma T, Tanaka M, Amano N, Watanabe A, Sakurai H, Yamamoto T, Yamanaka S, Hotta A (2015) Precise correction of the dystrophin gene in duchenne muscular dystrophy patient induced pluripotent stem cells by TALEN and CRISPR-Cas9. *Stem Cell Rep* 4:143–154
47. Ou Z, Niu X, He W, Chen Y, Song B, Xian Y, Fan D, Tang D, Sun X (2016) The combination of CRISPR/Cas9 and iPSC technologies in the gene therapy of human beta-thalassemia in mice. *Sci Rep* 6:32463
48. Chambliss AB, Chan DW (2016) Precision medicine: from pharmacogenomics to pharmacoproteomics. *Clin Proteomics* 13:25
49. Ashley EA (2016) Towards precision medicine. *Nat Rev Genet* 17:507–522
50. Hess GP, Fonseca E, Scott R, Fagerness J (2015) Pharmacogenomic and pharmacogenetic-guided therapy as a tool in precision medicine: current state and factors impacting acceptance by stakeholders. *Genet Res (Camb)* 97:e13
51. Gianni L, Eiermann W, Semiglazov V, Lluch A, Tjulandin S, Zambetti M, Moliterni A, Vazquez F, Byakhov MJ, Lichinitser M, Climent MA, Ciruelos E, Ojeda B, Mansutti M, Bozhok A, Magazzu D, Heinzmann D, Steinseifer J, Valagussa P, Baselga J (2014) Neoadjuvant and adjuvant trastuzumab in patients with HER2-positive locally advanced breast cancer (NOAH): follow-up of a randomised controlled superiority trial with a parallel HER2-negative cohort. *Lancet Oncol* 15:640–647
52. Hood L, Balling R, Auffray C (2012) Revolutionizing medicine in the 21st century through systems approaches. *Biotechnol J* 7:992–1001
53. Servant N, Romejon J, Gestraud P, La Rosa P, Lucotte G, Lair S, Bernard V, Zeitouni B, Coffin F, Jules-Clement G, Yvon F, Lermine A, Pouillet P, Liva S, Pook S, Popova T, Barette C, Prud'homme F, Dick JG, Kamal M, Le Tourneau C, Barillot E, Hupe P (2014) Bioinformatics for precision medicine in oncology: principles and application to the SHIVA clinical trial. *Front Genet* 5:152
54. Wolkenhauer O, Auffray C, Brass O, Clairambault J, Deutsch A, Drasdo D, Gervasio F, Preziosi L, Maini P, Marciniak-Czochra A, Kossow C, Kuepfer L, Rateitschak K, Ramis-Conde I, Ribba B, Schuppert A, Smallwood R, Stamatakos G, Winter F, Byrne H (2014) Enabling multiscale modeling in systems medicine. *Genome Med* 6:21
55. Tortolina L, Duffy DJ, Maffei M, Castagnino N, Carmody AM, Kolch W, Kholodenko BN, De Ambrosi C, Barla A, Biganzoli EM, Nencioni A, Patrone F, Ballestrero A, Zoppoli G, Verri A, Parodi S (2015) Advances in dynamic modeling of colorectal cancer signaling-network regions, a path toward targeted therapies. *Oncotarget* 6:5041–5058
56. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE (2015) Big data: astronomical or genomic? *PLoS Biol* 13:e1002195
57. Nault JC, Datta S, Imbeaud S, Franconi A, Mallet M, Couchy G, Letouze E, Pilati C, Verret B, Blanc JF, Balabaud C, Calderaro J, Laurent A, Letexier M, Bioulac-Sage P, Calvo F, Zucman-Rossi J (2015) Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. *Nat Genet* 47:1187–1193

What Can We Learn About Human Disease from the Nematode *C. elegans*?

Javier Apfeld and Scott Alper

Abstract

Numerous approaches have been taken in the hunt for human disease genes. The identification of such genes not only provides a great deal of information about the mechanism of disease development, but also provides potential avenues for better diagnosis and treatment. In this chapter, we review the use of the nonmammalian model organism *C. elegans* for the identification of human disease genes. Studies utilizing this relatively simple organism offer a good balance between the ability to recapitulate many aspects of human disease, while still offering an abundance of powerful cell biological, genetic, and genomic tools for disease gene discovery. *C. elegans* and other nonmammalian models have produced, and will continue to produce, key insights into human disease pathogenesis.

Key words *Caenorhabditis elegans*, Genetic screens, Genomic screens, RNAi, GFP

1 Introduction

The choice of model organism for study is a balance in trade-offs. While humans clearly are best in terms of mimicking human disease, there are practical and ethical limits to investigating disease in people. Other mammals, most notably mice, have proved very useful for modeling and studying human disease, but mice are limited in both how well they recapitulate some diseases and the ability to study them in rapid fashion. With the advent of tools like RNA interference (RNAi) and CRISPR/Cas9 genome editing, as well as more classical biochemical techniques, cell line studies have been very fruitful in identifying signaling pathways, for example, but are limited in that overall organismal physiology is generally not present in cell culture.

Non-mammalian model organisms such as the fruit fly *Drosophila melanogaster*, the zebrafish *Danio rerio*, and the nematode *Caenorhabditis elegans* serve as a happy medium [1–5], allowing for ease of study while still having the physiology present in a whole animal and the ability to recapitulate at least some aspects of human

disease. These and other model organisms have played key roles in human disease gene discovery. In the current review, we focus on the use of *C. elegans* as a nonmammalian model for human disease gene discovery. We first provide a brief introduction to *C. elegans* biology and the history of *C. elegans* research. Then we describe the key genetic and genomic techniques that have made *C. elegans* such a powerful research model. Using this background information, we illustrate two approaches that have been taken to identify human disease genes in *C. elegans*. In the first set of examples, we discuss how *C. elegans* disease models have been used for de novo discovery of human disease genes and pathways. In the second set of examples, we show how human disease genes have been engineered into *C. elegans* to develop models of human disease; these disease models have in turn been used to facilitate discovery of other genes that modulate that same human disease.

2 *C. elegans* Overview

“You have evolved from worm to man, but much within you is still worm.”
-Friedrich Nietzsche, Thus Spoke Zarathustra

2.1 What Is *C. elegans* Anyway?

Caenorhabditis elegans is a free living transparent nematode worm [6, 7] (Fig. 1). *C. elegans* starts out as an egg; when these eggs hatch, the nematodes pass through four larval stages before reaching adulthood. The *C. elegans* life cycle is relatively short, taking about 3 days for the animals to develop, and with an overall life span of about 2–3 weeks. Adults contain only 959 somatic nuclei and grow to be about a millimeter in length. Despite this small size, *C. elegans* has many of the organ systems present in more complex organisms,

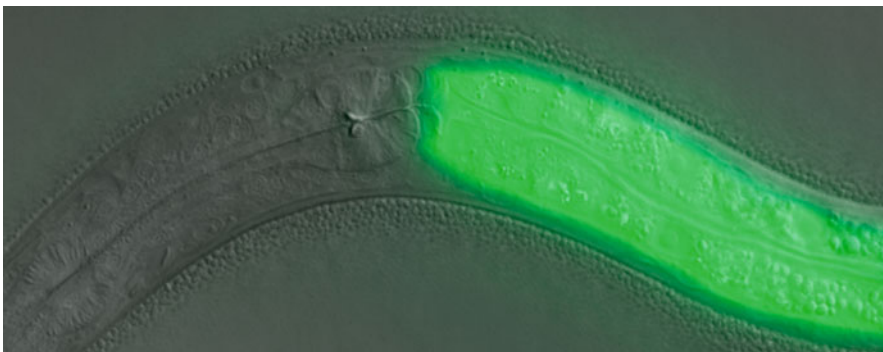


Fig. 1 Depicted is a *C. elegans* hermaphrodite carrying a *lys-7::gfp* transgene. In this animal, GFP expression is controlled by the gut-specific lysozyme-7 promoter. The image is an overlay of fluorescence and Nomarski images (images merged using Adobe Photoshop). Image adapted from Fig. 1 in [79]. Copyright © American Society for Microbiology, Molecular and Cellular Biology, 27, 2007, 5544–5553, doi:<https://doi.org/10.1128/MCB.02070-06>

including a digestive system, nervous system, musculature, and reproductive system. These small nematodes also exhibit complex behaviors. *C. elegans* will move toward things they like and away from things they do not like. The nematodes also eat, excrete, and mate.

C. elegans exists as either of two sexes, a hermaphrodite or a male. The existence of self-fertile hermaphrodites has great advantages for the study of development, because mutant stocks that would be unable to mate (such as paralyzed animals) are still able to self-fertilize. Moreover, healthy hermaphrodites produce hundreds of progeny, allowing the generation of large stocks quickly. When males are present, hermaphrodites can cross-fertilize. Thus, the presence of both sexes coupled with the relatively short life cycle allows for rapid genetic crosses.

C. elegans is only three cells in radius, with an outer epidermal layer, a middle muscle layer, and a central intestinal layer, with nervous system, reproductive system, and others tissues in between. The small size, transparent nature, and invariant cell lineage in *C. elegans* led to an unprecedented view of development in this animal. The full juvenile and adult cell lineages were reported more than 30 years ago [8, 9], and more recently, the entire wiring diagram of the nervous system has been determined [10]. In principle, if a cell is moved a few microns or a single neuronal connection is altered by some genetic manipulation, it should be possible to sort that out in *C. elegans*.

In the wild, *C. elegans* eats bacteria present in its environment [11]. In the laboratory, *C. elegans* typically is maintained on small petri dishes seeded with lawns of *E. coli* [12]. These bacteria are nonpathogenic and serve as a food source. Because of their small size, nematode manipulations are performed using a dissecting microscope. Individual nematodes can be moved from plate to plate using a small platinum wire “pick,” allowing investigators to isolate individual hermaphrodites for self-fertilization and the generation of large populations, or allowing investigators to set up crosses between the sexes. The small size of *C. elegans* means hundreds or thousands of animals can be maintained inexpensively on an individual dish. When the animals use up all the food, they will starve, and can be maintained as starved populations for months. For long-term storage of stocks, nematodes can be frozen and kept in frozen vials for decades at -80°C or in liquid nitrogen.

In summary, these little animals have many of the organs and exhibit many of the behaviors present in mammals. Moreover, they offer the ability to study diseases in the context of a whole, living, and intact organism, which is not possible in isolated cells. This has been particularly fruitful in the many diseases that affect behavior and the nervous system as described below. Roughly 30–60% of genes in *C. elegans* have orthologs or strong homologs in mammals [13, 14], suggesting that what is discovered about gene function in these small nematodes may be directly applicable to human development and disease.

2.2 Key Discoveries in *C. elegans*

The modern era of *C. elegans* research began over 50 years ago when Sydney Brenner first proposed using *C. elegans* to investigate developmental biology and neurobiology [15, 16]. Three of the notable discoveries that earned *C. elegans* researchers Nobel Prizes included the award to Sydney Brenner, Robert Horvitz, and John Sulston in 2002 for their discoveries related to development and the cell death machinery [17–19]; Andrew Fire and Craig Mello in 2006 for their discovery of RNA interference (RNAi) [20]; and Osamu Shimomura, Martin Chalfie, and Roger Tsien in 2008, for the discovery of Green Fluorescence Protein (GFP) [21, 22] and the demonstration that it could be a useful tool in other organisms including *C. elegans* [23]. Other key discoveries include the identification of microRNAs by Victor Ambros, Gary Ruvkun, and colleagues [24, 25]. For a more complete list of key discoveries, see [15].

3 The *C. elegans* Toolbox

The small size, rapid life cycle, and amazing genetic and genomic tools available have made *C. elegans* a premier model organism for many purposes. We outline some of these tools here.

3.1 Construction of Transgenic Nematodes

The *C. elegans* germ line initially develops as a multinucleate syncytium prior to membranes forming around each germ cell. Thus, DNA injected into the hermaphrodite gonad can be captured by numerous germ cells, making microinjection much easier than in other systems. DNA captured in this way will form extrachromosomal arrays that are semi-heritable [26, 27]. Selectable markers can then be used to maintain stable transgenic lines, and the DNA can be integrated into the genome if desired [28, 29]. In addition to direct microinjection, microparticle bombardment coupled with selection methods has been developed to generate stable nematode transgenic lines [30, 31]. More recently, sophisticated CRISPR/Cas9-based genome engineering strategies have enabled rapid and precise gene editing, thus facilitating the generation of animals bearing targeted point mutations, deletions, insertions and complex chromosomal rearrangements [32, 33].

The ease of *C. elegans* transgenic construction has served many purposes. Transgenic arrays can be used to restore gene function to “rescue” mutant phenotypes, greatly facilitating the cloning of mutated genes. Another common use for transgenic animals is the construction of GFP reporter strains. Promoter–GFP fusions can be used to determine where in the organism a particular gene is expressed. Protein–GFP fusions can be used for subcellular localization studies, and to quantify protein expression levels in live animals.

3.2 Genetic Tools and Forward Genetics in *C. elegans*

C. elegans is a diploid organism whose genome contains six chromosomes: five autosomes and one sex chromosome. XX animals are hermaphrodites; XO animals are males. The rapid lifecycle allows for quick genetic screens and crosses. Classical forward genetic screens used mutagens such as ethyl methanesulfonate (EMS) to randomly generate mutations in the nematode germ line [34–36]. F₁ hermaphrodite progeny that are heterozygous for these mutations can then be allowed to self-fertilize to isolate F₂ homozygous mutants of interest. If the homozygous mutant animals are self-fertile, they can be maintained as a homozygous stock. If the homozygous mutant animals are lethal or sterile, the screen can be engineered to recover heterozygous siblings to maintain the mutant stocks [37].

The ability to visualize *C. elegans* on a dissecting microscope or in more detail using a compound microscope equipped with differential interference contrast (DIC) optics allows for easy identification of mutant animals. Many classical mutants with visible phenotypes such as Unc (uncoordinated movement) or Dpy (dumpy shaped animals) were isolated by mutagenesis and visual screening for morphological or behavioral phenotypes [38]. More recently, screens have been performed for worms with altered levels or location of GFP expression, altered movement, or altered learning, and almost anything else *C. elegans* researchers can imagine. There are numerous mapping strategies to determine the identity of the mutant genes ranging from crosses with strains carrying known genetic markers, SNP mapping strains, strains carrying deletion chromosomes, or balancer chromosomes [7, 39]. Once the mutation is mapped to a region where a candidate gene is found, the wild type copy of the locus can be injected into animals in an attempt to rescue the mutant phenotype. Alternatively or additionally, RNAi can be delivered to the animals in an attempt to phenocopy the mutant phenotype. The candidate locus also can be sequenced to identify mutations, although more and more frequently whole genome sequencing is being used to identify the causative mutation [40, 41]. To simplify mapping and mutation identification, transposon-mediated mutagenesis is also an option in *C. elegans* [35, 42].

In addition to classical forward genetic screens, many researchers have used modifier screens with great success [34–36]. In this case, researchers start with a strain carrying a mutation that induces a phenotype and then mutagenize the animals to isolate mutant animals harboring suppressor or enhancer mutations. For example, one could start with a mildly uncoordinated animal, mutagenize, and screen visually using the dissecting microscope for suppressors that restore normal movement. These modifier mutations can then be genetically separated from the original mutation to determine if the modifier mutation has a phenotype on its own.

The ability to perform rapid genetic crosses also makes *C. elegans* an excellent system to perform genetic epistasis studies to place novel mutations in known genetic pathways [36].

3.3 Genomic Tools and Reverse Genetics in *C. elegans*

The discovery of RNAi opened up a whole new world for researchers in all fields including investigators studying *C. elegans*. Because there is no interferon response in *C. elegans*, long dsRNAs are not toxic to the nematode. Thus, long dsRNAs rather than siRNAs can be delivered to *C. elegans* with a concomitant increase in efficiency and specificity of knockdown. *C. elegans* RNAi screens generally do not suffer from the off-target effects that have plagued mammalian screens. The method of dsRNA delivery in *C. elegans* is also unique. Andy Fire and colleagues demonstrated that *E. coli* engineered to express dsRNA can be fed to *C. elegans*, resulting in knockdown of the target gene [43]. Taking advantage of this technique of RNAi feeding, the Ahringer and Vidal labs have generated two genomic RNAi bacterial feeding libraries that cover most of the *C. elegans* genome [44, 45]; each bacterial strain enables the specific RNAi knockdown of a single gene, allowing for rapid and simple genome-wide screening. In these genomic RNAi screens, one simply feeds the bacteria to the nematodes, one bacterial strain at a time, and monitors for the occurrence of the phenotype of interest. Additionally, mutations that enhance RNAi-mediated knockdown have been identified and used to increase the sensitivity of these RNAi screens [46, 47].

While RNAi is an invaluable tool, ultimately it is important to be able to monitor the effect of mutation of genes of interest. Unlike RNAi gene knockdowns, mutations allow for less heterogeneous effects. Mutations also can cause unique effects in gene function, such as gain of function or dominant-negative effects. Several labs that make up the *C. elegans* knockout consortia have been isolating thousands of knockout mutations available to the community of *C. elegans* researchers [48, 49]. Likewise, the *Caenorhabditis* Genetics Center (CGC) is a stock center that provides ready access to these mutations and the myriad of other mutations that have been isolated and shared by the *C. elegans* research community. More recent targeted transposon insertion [50], and CRISPR/Cas9 genome editing [32, 33] approaches have further enhanced the ability to perform reverse genetics in the nematode by enabling the introduction of almost any change in any gene in the genome.

4 Identifying Novel Human Disease Genes in *C. elegans*

In the next two sections, we outline several representative examples of human disease gene identification in *C. elegans*. We apologize to researchers whose work could not be included due to space

limitations. Rather than aiming to be comprehensive, our goal is to be illustrative. These specific examples have been chosen to illustrate (1) the advantages of the techniques available in *C. elegans* to facilitate disease gene discovery and (2) some of the follow-up studies in mammals that have been performed. For de novo disease gene discovery, we outline various genetic and genomic screens for regulators of innate immunity, obesity, and aging (Subheadings 4.1–4.3). For human disease model studies in *C. elegans*, we outline the investigation of various neurodegenerative diseases (Subheading 5).

4.1 Innate Immunity

Infectious and inflammatory diseases are among the leading causes of death throughout the world. Infectious diseases account for five of the top ten causes of death in the developing world [51]. In developed countries, the top three leading causes of death are heart disease, cancer, and COPD [52]. A key factor common to these three diseases is chronic inflammation [53–56]. This illustrates the importance of proper regulation of innate immunity and inflammation. While a robust innate immune response is essential in our pathogen-rich world, this response must be tightly regulated to prevent inflammatory disease. The identification of genes that regulate innate immunity has led to the identification of numerous genes that affect infectious or inflammatory disease [53, 54, 57–62].

C. elegans has emerged as a key model system for the discovery of innate immune genes [63–65]. For decades, *C. elegans* researchers cultured *C. elegans* on petri dishes containing lawns of non-pathogenic *E. coli*. However, Ausubel and colleagues discovered that by simply replacing this *E. coli* lawn with any of a number of human pathogens, the bacteria would infect and kill *C. elegans* [66–68]. Since then, pathogenesis models have been developed for Gram negative and positive bacteria, fungi, and viruses [69–72]. *C. elegans* lacks migratory immune cells and does not have an adaptive immune response. The nematode innate immune response is composed of the production of antimicrobial peptides and compounds that fight infection [73]. Importantly, the induction of antimicrobial production in the presence of pathogens is mediated by conserved signaling pathways including MAP kinase cascades [74]. However, there also are differences, most notably the absence of an NFκB homolog in *C. elegans*. Many investigators have now used *C. elegans* to study host-pathogen interactions.

Irazoqui and colleagues took a variety of approaches to identify a novel innate immunity regulatory pathway conserved in *C. elegans* and mammals. They first monitored changes in *C. elegans* gene expression induced by infection with the Gram positive bacterial pathogen *S. aureus* [75]. They then used computational analysis of these data to determine that the *C. elegans* HLH-30 transcription factor (mammalian ortholog TFEB) target DNA sequence was overrepresented in the promoters of the genes whose expression

was induced by *S. aureus*. To test if HLH-30 was involved in this response, they generated HLH-30-GFP transgenic nematodes and found that while HLH-30-GFP was present in both the nucleus and cytoplasm in uninfected worms, all the HLH-30-GFP was present in the nucleus following infection [76]. They then used RNAseq to monitor *S. aureus*-induced gene expression changes in wild type and *hlh-30* mutant animals and discovered that much of the *S. aureus*-induced gene expression was dependent on the function of HLH-30; moreover, both HLH-30 and its target genes were required for full resistance to *S. aureus* [76]. This approach illustrates several advantages of the nematode system, including the ease of generating transgenic animals, localization of GFP fusions in the transparent nematode, the availability of a deletion mutant in *hlh-30*, and the availability of bacteria to deliver *hlh-30* dsRNA. Moreover, the identification of HLH-30/TFEB as a key innate immunity regulator was validated in mammalian cells. *S. aureus* infection in mammalian cell culture leads to redistribution of TFEB into the nucleus, and inhibition of TFEB weakens the *S. aureus*-induced pro-inflammatory response [75]. Using knockout mice, other investigators have independently shown that TFEB affects innate immunity in mammals [77], providing further evidence of the validity of the *C. elegans* studies.

In a follow-up to these studies, Irazoqui and colleagues used a targeted RNAi screen in which they inhibited most of the kinases and phosphatases in the nematode genome. This targeted RNAi screening approach led to the identification a PLC-PKD-TFEB pathway regulating the nematode innate immune response [78]. They took advantage of the ease of nematode genetics to order the various genes into a pathway, and then went on to show that this signaling pathway functioned similarly in mouse macrophages [78]. This highlights the importance of the *C. elegans* approach. Similar RNAi screens in mammals would have been significantly more cumbersome and expensive, and it would have been much more complicated to perform the genetic epistasis studies to determine how these genes functioned in an ordered pathway. However, once these details were worked out in *C. elegans*, the confirmatory cell culture RNAi studies were much more straightforward.

We have used a slightly different strategy with similar results: using *C. elegans* as a rapid screening tool with follow-up studies in mammalian cells and mice. We used comparative genomics RNAi screens in *C. elegans* and mouse macrophages to identify innate immunity regulators, subsequently used *C. elegans* infection models to obtain in vivo validation of these RNAi data, and then used knockout mice to determine the effect of these genes in mammalian disease. We used the ease of generating nematode transgenics to generate 14 different antimicrobial-GFP reporter strains [79]. GFP expression in these lines could be monitored using fluorescence

microscopy or by using the COPAS Biosort, a flow cytometer for *C. elegans* [80]. A key feature of the COPAS Biosort is that it can analyze nematodes in 96-well format, allowing for high-throughput screens. We used bacterial feeding RNAi to inhibit known innate immunity regulators in *C. elegans*, and found several antimicrobial-GFP reporters whose expression was regulated by these known pathways. This formed the basis for a genomic RNAi screen in which we screened for changes in antimicrobial-GFP levels in the presence of *E. coli*. To determine if the genes identified could regulate innate immunity in mammals, siRNAs targeting the mouse orthologs of these genes were delivered into mouse macrophage cell lines and the cytokine response induced by lipopolysaccharide (LPS) was monitored. Remarkably, 30–40% of the genes identified in *C. elegans* had an RNAi-induced defect in the innate immune response in mouse macrophages [81–83]. The ready availability of existing *C. elegans* knockouts allowed us to rapidly obtain in vivo confirmation that these genes affected host defense. We found that nine of ten *C. elegans* knockouts tested had altered survival in the presence of the nematode and human pathogen *P. aeruginosa* [81–83]. Armed with the RNAi data in *C. elegans* and mouse macrophages, and *C. elegans* knockout data, we then tested four different mouse knockout lines and found that three of the four knockout mice exhibited an altered innate immune response when challenged with LPS ([83, 84] and unpublished). Thus, our comparative genomics approach is an efficient method for finding novel innate immunity regulators.

There are several things worth noting about this approach. First, one of the complications of RNAi screens in mammalian cells is the high degree of false-positives due to off-target effects [85]. This is likely not a problem in *C. elegans* because of the use of long dsRNAs. Moreover, the screens in *C. elegans* and macrophages involved different methods of dsRNA delivery, different innate immune stimuli, and different immunological readouts. It seems highly unlikely that such different systems would coincidentally report similar results. Plus, the ability to obtain so many nematode mutants relatively rapidly and cheaply for in vivo validation would just not be plausible in mice. By the time these genes had passed all these tests, the efficiency of validating them in vivo in mice was very high. Mammalian follow-up studies focused on genes identified in these screens have led to the investigation of two pathways that regulate the maintenance but not the activation phase of innate immunity [84, 86, 87].

4.2 Obesity

Obesity has become an epidemic in developed countries; more than 1/3 of adults in the USA are now obese [88]. Obesity is among the leading causes of preventable death and also affects many comorbidities such as type 2 diabetes [89]. The excess fat accumulation in obesity is caused by both genetic and environmental factors

[90]. The ability to monitor fat accumulation in *C. elegans* coupled with the ease of RNAi screening in the nematode has led to a number of studies that identified genes that control fat accumulation [91–93]. In one study, McKay et al. [94] demonstrated that RNAi-mediated inhibition of genes known to affect fat accumulation in mammals, including SREBP and C/EBP homologs, led to arrested *C. elegans* development. Moreover, these animals did not accumulate fat [94], as assayed using Sudan Black or Nile Red staining. The authors reasoned that inhibition of other genes that affect fat production would likewise arrest larval development and would be lethal. The investigators used RNAi to inhibit 80 genes known to be larval-lethal when inhibited, and discovered that ten gene inhibitions affected fat accumulation. They then used RNAi to verify that these genes affected mammalian cells as well [94]. Ashrafi et al. [95] used genome-wide RNAi screens followed by Nile Red staining to identify the full complement of genes that alter fat accumulation in *C. elegans*; these investigators identified 305 gene inactivations that reduced fat accumulation and 112 gene inactivations that increased fat accumulation. In another approach, a GFP reporter that localized to fat droplets was used as a screening tool to identify RNAi treatments that altered fat accumulation [96]. All these studies, and many others, demonstrate the ease of RNAi screening in *C. elegans* coupled with the effective readout tools available to study different diseases in a transparent organism.

4.3 Aging

The study of aging in *C. elegans* is unusual in that prior to these investigations, most researchers would not have even considered aging a disease that could be investigated and manipulated genetically. Thus, not only have *C. elegans* studies of aging been fruitful for finding potential human disease genes, but these studies also established that aging was a phenomenon that could be studied genetically in the first place.

As we grow older, we become increasingly frail and eventually die. Age is a major risk factor for a wide variety of diseases. These include almost all of the major neurodegenerative diseases, such as Alzheimer's disease and Parkinson's disease, as well as cardiovascular disease, metabolic disease, and many cancers. Until recently, aging was not considered a genetically tractable phenomenon and instead was thought to result from the unregulated accumulation of all sorts of errors that together lead to the decay in function and death of the organism. As a result, our understanding of the mechanisms of aging was very poor. Over the last 25 years, however, our understanding of aging has been transformed by pioneering studies in *C. elegans*. Powerful genetics coupled with a relatively short life span of 20 days make *C. elegans* an excellent system to study aging. Its short life span makes it possible to conduct experiments that just are not practical in mice (mean life span of 2 years) or humans (mean life span of 80 years). In addition, its simple and

inexpensive ease of manipulation makes it possible to assay the life span of hundreds or even thousands of worms. These studies have shown that aging is a regulated phenomenon that can be studied with the tools of molecular biology and genetics, and that many of the genes that regulate aging in nematodes also regulate aging in other organisms, including *Drosophila*, mice, and possibly humans.

The first forward-genetic screen for long-lived *C. elegans* mutants was conducted in the 1980s by Michael Klass [97]. This elegant genetic screen surmounted several technical challenges specific to *C. elegans* aging studies. Nematodes produce hundreds of progeny, and thus, parents will rapidly be lost among their progeny as they grow on small petri dishes. To measure the life span of a population of worms, one has to separate each worm from its progeny, typically by daily transfer to new petri plates until reproduction ends. This is a very cumbersome process. Moreover, once a mutant worm is deemed long-lived, one needs to obtain progeny to maintain a mutant line that can be studied; however, old worms are no longer fertile. Klass overcame these two challenges using a known temperature-sensitive spermatogenesis mutation. After mutagenesis, F₂ animals were each transferred singly to new “master” plates where they reproduced at the lower permissive temperature. Some of the F₃ progeny were grown at a high “restrictive” temperature, where they developed into animals that could not self-fertilize. Klass determined the life span of thousands of such cohorts to identify eight long-lived mutants. He reisolated these mutants from their respective master plates that were maintained at the permissive temperature, since their siblings had the same mutations. Three of these mutations were subsequently mapped and shown to be in the same genetic locus, named *age-1* [98, 99]. Remarkably, *age-1* mutant animals lived more than twice as long as wild-type control animals. These studies showed that mutations in a single gene could have a dramatic effect on the life span of a multicellular organism.

A few years later, Cynthia Kenyon’s laboratory discovered that mutations in another gene, *daf-2*, could more than double *C. elegans* life span; moreover, the aged *daf-2* mutant animals remained youthful in appearance and mobility, even when all wild-type control animals had died [100]. The *daf-2* mutation was previously known to also affect the developmental decision to form dauer larvae [101]. Under unfavorable growth conditions of high-temperature, low food, and high population density, *C. elegans* develops into developmentally arrested, stress-resistant, nonfeeding dauer larvae; dauers can resume development into fertile adults once they encounter a more favorable environment [102]. *daf-2* mutant animals were known to inappropriately form dauer larvae at high temperature, but in an otherwise favorable growth environment. Kenyon and colleagues showed that at a low temperature where these mutant animals did not form dauers,

they instead developed into fertile adults that were long-lived. A few years earlier, the Riddle lab [101, 103, 104] had performed several genetic screens and assembled a genetic pathway for the regulation of dauer formation. Kenyon and colleagues took advantage of this knowledge and asked whether a similar regulatory pathway existed for life span. They found that *daf-16*, a gene required for *daf-2* mutant animals to form dauer larvae, is also necessary for the increased life span of *daf-2* mutant adults [100]. Taken together, these findings demonstrated that aging is subject to regulation.

Subsequent studies have shown that *daf-2*, *age-1*, and *daf-16* are all part of a conserved insulin/IGF1 signaling pathway: *daf-2* encodes the worm's only ortholog of the human insulin and IGF1 receptor tyrosine kinases; *age-1* encodes a phosphoinositide 3-kinase (PI-3 kinase); and *daf-16* encodes a FOXO transcription factor that is negatively regulated by the *age-1* effector kinases AKT-1 and AKT-2 [105, 106]. These genes are part of a well-conserved signaling pathway, raising the question of whether insulin/IGF1 signaling likewise regulated life span in other organisms. Subsequent studies in the fruit fly *Drosophila melanogaster* [106–110] and mice [111, 112] showed that manipulation of the insulin/IGF1 signaling pathway can increase life span in fruit flies and mice. While these follow-up mouse studies were critical to demonstrate that these pathways were conserved in mammals, these studies highlight the practicality of forward genetic screens for life span in *C. elegans*, which would be a much more challenging in mice.

These remarkable studies prompted the question of whether similar mechanisms may regulate aging in humans [113, 114]. Several candidate-based and unbiased association studies have since identified variants in the *daf-16* ortholog *FOXO3A* that are associated with exceptional longevity in humans from multiple ethnic origins [115–124]. In addition, mutations in the IGF1 receptor gene that cause diminished IGF1 signaling were found to be more prevalent in a cohort of Ashkenazi Jewish centenarians, compared to control individuals that do not exhibit exceptional longevity [125, 126]. Taken together, these findings suggest that differences in human life span may result, at least in part, from the normal variation in signaling by the IGF1 receptor and its transcriptional effector *FOXO3A*.

Since the discovery of the regulation of life span by insulin/IGF-1 signaling, the study of aging in *C. elegans* has exploded, leading to the discovery of hundreds of genes that affect life span. These life span-determining genes have been identified by a combination of forward and reverse-genetic approaches. One of the most fruitful approaches has been to determine the effect of each gene on life span by systematically knocking down each gene in the genome using RNAi. To date, three such genome-wide RNAi

screens have been completed [127–129]. In addition, a genome-wide RNAi screen was performed to identify genes whose knock-down shortens life span in *daf-2* mutant animals [130], as well as numerous more targeted screens [131, 132]. It likely will take many years until all these discoveries are replicated in mammalian systems, but investigators are already tackling the question of whether aging may be “druggable,” potentially leading to an extension of “health span” and life span, and a delay in the onset of many age-related diseases [133, 134].

5 Modeling Human Diseases in *C. elegans*

In contrast to the above approaches which involve screening de novo in *C. elegans* for genes that alter a phenotype that is involved in human disease, an alternate approach has been to artificially engineer the human disease into *C. elegans*, typically by expressing the human disease gene in the nematode. Animals engineered to exhibit the human disease are then used as tools to screen for suppressors or enhancers of the disease phenotype with the goal of finding additional gene targets that affect the disease in humans. While there are many examples of this approach, they are perhaps best exemplified by the study of neurodegenerative disorders in *C. elegans*, as outlined below.

5.1 Poly-Glutamine Repeat Diseases

Trinucleotide repeat diseases are typically neurodegenerative or neuromuscular disorders caused by inheritance of a trinucleotide repeat (often greater than 30 repeats in length) in particular genes [135–139]. These trinucleotide repeats are formed by the expansion of unstable shorter triplet repeats present in the genome [135–139]. Some of the most studied triplet repeat disorders are caused by expansion of CAG repeats. These are the poly-Glutamine (polyQ) repeat diseases, which include Huntington’s disease, spinocerebellar ataxias, and many others. Key questions about such disorders include how these unstable repeats expand in the genome, why there is apparently a threshold length for the repeat beyond which disease occurs, and how to develop possible treatments.

Expression of polyQ repeat proteins in *C. elegans* muscle [140] or neurons [141, 142] recapitulates some aspects of human polyQ disease. In particular, some of these authors and others have found a similar threshold for the number of repeats that cause disease. Expression of roughly 35–40 repeats of polyQ-YFP were required to induce polyQ-protein aggregation and resulting muscle or neuronal dysfunction. The ability to monitor YFP-tagged polyQ-protein aggregation in this transparent organism allowed for straightforward modifier screens to monitor polyQ-induced aggregation or dysfunction. For example, Nollen et al. [143] used a genomic RNAi screen to identify 186 genes whose inhibition led

to increased or earlier onset aggregation of Q35-YFP (polyQ protein with 35 Q repeats). These genes fell into five broad functional categories, including regulation of RNA metabolism, protein synthesis, protein folding, protein degradation, and protein trafficking. Similarly, candidate based-approaches have been used to identify modifiers of polyQ aggregation in *C. elegans*. For example, overexpression of the *C. elegans* homolog of the torsin gene suppressed polyQ aggregation [144]. Likewise, overexpression of ubiquitin suppressed polyQ-induced toxicity in *C. elegans* and mammalian cells while inhibition of ubiquitin expression induced the opposite effect [145]. The ease of such genetic and genomic studies in *C. elegans* coupled with the ability to monitor fluorescently tagged polyQ proteins in this transparent organism has made such studies very straightforward and powerful.

5.2 Alzheimer's Disease

Alzheimer's disease is the sixth leading cause of death in the USA, affecting more than 5 million people in the USA and more than 35 million people worldwide [146, 147]. As is the case for most age-dependent diseases, the incidence of Alzheimer's disease is expected to increase in the future. Despite intensive study, much about Alzheimer's disease remains a mystery, and no effective treatments have been developed. Much of the research focus centers on trying to understand the aggregation of proteins such as Tau or beta amyloid and the resulting effects on neurological function [146, 147].

Several investigators have used overexpression of wild type or mutant Tau as a model for tauopathy in *C. elegans* [148, 149]. Kraemer and colleagues [150] expressed wild type or mutant Tau in all nematode neurons; they observed that Tau aggregated in these animals and that Tau overexpression led to a moderate uncoordinated phenotype. They used this model as the basis for a genome-wide RNAi screen for enhancers of this uncoordinated phenotype [151]. The genes and pathways identified in this screen as potential modifiers of Tau-induced pathology were very similar to those identified in *Drosophila* screens, suggesting that they may be conserved regulators that might play a role in tauopathies and Alzheimer's disease [152]. In addition to their genomic RNAi screen, the investigators performed a forward genetic screen to identify mutations that suppress the Tau-induced uncoordinated phenotype. In this genetic screen, they identified mutations in *sut-2*, which suppressed the Tau aggregation, uncoordinated, and neurodegenerative phenotypes induced by Tau overexpression in *C. elegans* [153]. Moreover, overexpression of *sut-2* in nematodes exacerbated Tau-induced neurotoxicity, the opposite of the RNAi-induced phenotype [154]. The role of SUT-2 was not unique to *C. elegans*. Follow-up studies in mammalian cells demonstrated that (1) Tau overexpression increased expression of the mammalian homolog MSUT2, (2) MSUT2 RNAi in mammalian cells

diminished aggregation of insoluble Tau, and (3) there is less MSUT2 present in the brain in autopsy samples from Alzheimer's disease patients [154]. Thus, these genomic and genetic modifier screens in *C. elegans* successfully identified key genes to investigate in the human disease.

Studies in *C. elegans* relevant to Alzheimer's disease are not limited to the investigation of Tau. For example, beta-amyloid-expressing models of disease have also been engineered in *C. elegans* [155–160]. Likewise, investigation of the nematode homologs of Presenilin 1 and 2, mutations in which cause early-onset familial Alzheimer's disease [161–163], led to the discovery that nematode and human Presenilin 1 regulates Notch signaling [164–166]. As an illustration of the power of genetic screens in *C. elegans*, our lab conducted a sensitized forward genetic screen in *C. elegans* to identify genes that function with the Presenilins. In this screen, we identified mutations in two novel genes (*aph-1* and *pen-2*) that enhanced the phenotype induced by mutation of *sel-12* (Presenilin) [167]. *aph-1* also was identified in a *C. elegans* genetic screen for enhancers of Notch signaling [168]. These genes were later shown to be part of the evolutionarily conserved γ -secretase protease complex, where they regulate the maturation of Presenilin [169], the catalytic component of this complex. This complex is involved in the proteolytic maturation or degradation of many transmembrane proteins, including the Amyloid Precursor Protein (APP), which is important in Alzheimer's disease pathogenesis, and the Notch receptor.

5.3 Parkinson's Disease

Parkinson's disease is second only to Alzheimer's disease as the most common neurodegenerative disease. Like Alzheimer's disease, Parkinson's disease usually, but not exclusively, is an age-dependent disease, with an incidence of roughly 1% in people over 65 rising to an incidence of 5% by age 85 [170–172]. The primary cause of Parkinson's disease is a loss of dopaminergic neurons in the substantia nigra region of the brain. This results in the neurological symptoms that are a hallmark of the disease, including tremor of the hands, legs, limbs, and jaw, muscle rigidity of the limbs and trunk, bradykinesia, and postural instability. A key histological feature of patients with Parkinson's disease is the accumulation of Lewy Bodies in the brain.

A number of genomic and candidate gene-based RNAi screens have been performed in *C. elegans* models of Parkinson's disease. These models have focused on overexpression of α -Synuclein, a key candidate Parkinson's disease protein. α -Synuclein is the main component of Lewy Bodies. It is overexpressed and often misexpressed in the brain of Parkinson's disease patients, and mutations in α -Synuclein have been identified in some patients [173, 174]. α -Synuclein is not present in *C. elegans*. Nematode researchers have taken advantage of this to overexpress α -Synuclein and screen for

genes that affect α -Synuclein aggregation or cell function [175]. In two studies, YFP or GFP-tagged human α -Synuclein was expressed in nematode muscle using cell-type specific promoters. Aggregated α -Synuclein was monitored by the appearance of punctate fluorescent structures, and either a genomic RNAi screen [176] or an RNAi screen of 900 priority candidate genes (based on various bioinformatics approaches) [177] led to the discovery of numerous genes that affect α -Synuclein aggregation. Many of these genes, in turn, were found to serve a neuroprotective function.

RNAi screens focusing on neurons in *C. elegans* are more challenging because nematode neurons are somewhat resistant to RNAi. Thus, to study the effects of α -Synuclein expressed in neurons, Kuwahara et al. [178] took advantage of a mutation, *eri-1*, that enhances RNAi in *C. elegans*. They expressed human α -Synuclein in all nematode neurons in a strain carrying this *eri-1* mutation. Under these conditions, there was little gross effect on the animals. They then performed an enhancer RNAi screen targeting 1673 prioritized candidate genes (genes known to affect the nervous system) to identify RNAi treatments that induced a visible phenotype such as uncoordinated movement or growth retardation. Ten candidate genes passed their screening criteria; four of these genes functioned in the endocytic machinery, implicating endocytosis in the pathogenesis of α -Synuclein.

6 Conclusion

The choice of models to investigate human disease is often a trade-off between how well the model mimics the human condition and how easy it is to manipulate the system. Invertebrate models such as *C. elegans* and *D. melanogaster* have been invaluable for the study of development, signaling pathways, and many other aspects of biology. In this chapter, we have outlined several examples that illustrate the ease of such *C. elegans* studies. Some of the features that have rendered *C. elegans* such a powerful research organism include the ease of genetics (forward genetic screening, transgenic animal construction, mutation mapping), cell biology (using GFP in a transparent organism with a fully described and invariant cell lineage), genomics (RNAi and other techniques), modifier screens (enhancement and suppression), and the ability to mimic many human diseases. We also have highlighted how more and more frequently, follow-up studies in mammals have validated these nematode findings. The tools and ease of use of *C. elegans* and other “simple” model organisms continues to make them invaluable for research, and these organisms will continue to play an important role in our understanding of human disease and human disease gene discovery in the future.

Acknowledgments

This work was supported by National Institutes of Health Grant 1R01ES025161 to S.A.

References

- Aitman TJ, Boone C, Churchill GA, Hengartner MO, Mackay TF, Stemple DL (2011) The future of model organisms in human disease research. *Nat Rev Genet* 12(8):575–582
- Lieschke GJ, Currie PD (2007) Animal models of human disease: zebrafish swim into view. *Nat Rev Genet* 8(5):353–367
- Lopez Hernandez Y, Yero D, Pinos-Rodriguez JM, Gibert I (2015) Animals devoid of pulmonary system as infection models in the study of lung bacterial pathogens. *Front Microbiol* 6:38
- Markaki M, Tavernarakis N (2010) Modeling human diseases in *Caenorhabditis elegans*. *Biotechnol J* 5(12):1261–1276
- Pandey UB, Nichols CD (2011) Human disease models in *Drosophila melanogaster* and the role of the fly in therapeutic drug discovery. *Pharmacol Rev* 63(2):411–436
- Riddle DL, Blumenthal T, Meyer BJ, Priess JR (eds) (1997) *C. elegans* II, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY)
- Wood WB (1988) The nematode *Caenorhabditis elegans*. Cold Spring Harbor Laboratory Press, New York
- Sulston JE, Horvitz HR (1977) Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev Biol* 56(1):110–156
- Sulston JE, Schierenberg E, White JG, Thomson JN (1983) The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev Biol* 100(1):64–119
- Varshney LR, Chen BL, Paniagua E, Hall DH, Chklovskii DB (2011) Structural properties of the *Caenorhabditis elegans* neuronal network. *PLoS Comput Biol* 7(2):e1001066
- Samuel BS, Rowedder H, Braendle C, Felix MA, Ruvkun G (2016) *Caenorhabditis elegans* responses to bacteria from its natural habitats. *Proc Natl Acad Sci U S A* 113(27):E3941–E3949
- Lewis JA, Fleming JT (1995) Basic culture methods. *Methods Cell Biol* 48:3–29
- Shaye DD, Greenwald I (2011) OrthoList: a compendium of *C. elegans* genes with human orthologs. *PLoS One* 6(5):e20085
- Sonnhammer EL, Durbin R (1997) Analysis of protein domain families in *Caenorhabditis elegans*. *Genomics* 46(2):200–216
- Corsi AK, Wightman B, Chalfie M (2015) A transparent window into biology: a primer on *Caenorhabditis elegans*. *Genetics* 200(2):387–407
- Strange K (2006) An overview of *C. elegans* biology. *Methods Mol Biol* 351:1–11
- Ellis HM, Horvitz HR (1986) Genetic control of programmed cell death in the nematode *C. elegans*. *Cell* 44(6):817–829
- Check E (2002) Worm cast in starring role for Nobel prize. *Nature* 419(6907):548–549
- Marx J (2002) Nobel prize in physiology or medicine. Tiny worm takes a star turn. *Science* 298(5593):526
- Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391(6669):806–811
- Roda A (2010) Discovery and development of the green fluorescent protein, GFP: the 2008 Nobel prize. *Anal Bioanal Chem* 396(5):1619–1622
- Kricka LJ, Stanley PE (2009) Scientists awarded Nobel prize for work with GFP. *Luminescence* 24(1):1
- Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC (1994) Green fluorescent protein as a marker for gene expression. *Science* 263(5148):802–805
- Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75(5):843–854
- Wightman B, Ha I, Ruvkun G (1993) Post-transcriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* 75(5):855–862
- Mello C, Fire A (1995) DNA transformation. *Methods Cell Biol* 48:451–482
- Stinchcomb DT, Shaw JE, Carr SH, Hirsh D (1985) Extrachromosomal DNA transformation of *Caenorhabditis elegans*. *Mol Cell Biol* 5(12):3484–3496

28. Mello CC, Kramer JM, Stinchcomb D, Ambros V (1991) Efficient gene transfer in *C. elegans*: extrachromosomal maintenance and integration of transforming sequences. *EMBO J* 10(12):3959–3970
29. Fire A (1986) Integrative transformation of *Caenorhabditis elegans*. *EMBO J* 5(10):2673–2680
30. Praitis V, Casey E, Collar D, Austin J (2001) Creation of low-copy integrated transgenic lines in *Caenorhabditis elegans*. *Genetics* 157(3):1217–1226
31. Schweinsberg PJ, Grant BD (2013) *C. elegans* gene transformation by microparticle bombardment. *WormBook: the online review of C elegans biology*, Pasadena (CA), pp 1–10
32. Dickinson DJ, Goldstein B (2016) CRISPR-based methods for *Caenorhabditis elegans* genome engineering. *Genetics* 202(3):885–901
33. Kim HM, Colaiacovo MP (2016) CRISPR-Cas9-guided genome engineering in *C. elegans*. *Curr Protoc Mol Biol* 115:31.37.1–31.37.18
34. Anderson P (1995) Mutagenesis. *Methods Cell Biol* 48:31–58
35. Kutscher LM, Shaham S (2014) Forward and reverse mutagenesis in *C. elegans*. *WormBook: the online review of C elegans biology*, Pasadena (CA), pp 1–26
36. Wang Z, Sherwood DR (2011) Dissection of genetic pathways in *C. elegans*. *Methods Cell Biol* 106:113–157
37. Fay DS (2013) Classical genetic methods. *WormBook: the online review of C elegans biology*, Pasadena (CA), pp 1–58
38. Brenner S (1974) The genetics of *Caenorhabditis elegans*. *Genetics* 77(1):71–94
39. Williams BD (1995) Genetic mapping with polymorphic sequence-tagged sites. *Methods Cell Biol* 48:81–96
40. PJ H (2014) Whole genome sequencing and the transformation of *C. elegans* forward genetics. *Methods* 68(3):437–440
41. Hobert O (2010) The impact of whole genome sequencing on model system genetics: get ready for the ride. *Genetics* 184(2):317–319
42. Bessereau JL (2006) Transposons in *C. elegans*. *WormBook: the online review of C elegans biology*, Pasadena (CA), pp 1–13
43. Timmons L, Fire A (1998) Specific interference by ingested dsRNA. *Nature* 395(6705):854
44. Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, Ahringer J (2000) Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* 408(6810):325–330
45. Rual JF, Ceron J, Koreth J, Hao T, Nicot AS, Hirozane-Kishikawa T, Vandenhaute J, Orkin SH, Hill DE, van den Heuvel S, Vidal M (2004) Toward improving *Caenorhabditis elegans* phenome mapping with an ORFeome-based RNAi library. *Genome Res* 14(10B):2162–2168
46. Kennedy S, Wang D, Ruvkun G (2004) A conserved siRNA-degrading RNase negatively regulates RNA interference in *C. elegans*. *Nature* 427(6975):645–649
47. Simmer F, Tijsterman M, Parrish S, Koushika SP, Nonet ML, Fire A, Ahringer J, Plasterk RH (2002) Loss of the putative RNA-directed RNA polymerase RRF-3 makes *C. elegans* hypersensitive to RNAi. *Curr Biol* 12(15):1317–1319
48. Consortium CeDM (2012) Large-scale screening for targeted knockouts in the *Caenorhabditis elegans* genome. *G3* 2(11):1415–1425
49. Mitani S (2009) Nematode, an experimental animal in the national BioResource project. *Exp Anim* 58(4):351–356
50. Frokjaer-Jensen C (2015) Transposon-assisted genetic engineering with mos1-mediated single-copy insertion (MosSCI). *Methods Mol Biol* 1327:49–58
51. World Health Statistics 2014 (2015)
52. Detailed Tables for the National Vital Statistics Report (NVSR) (2015) “Deaths: Final Data for 2013”. vol 64
53. Chaudhuri N, Dower SK, Whyte MK, Sabroe I (2005) Toll-like receptors and chronic lung disease. *Clin Sci* 109(2):125–133
54. Cook DN, Pisetsky DS, Schwartz DA (2004) Toll-like receptors in the pathogenesis of human disease. *Nat Immunol* 5(10):975–979
55. Grivennikov SI, Greten FR, Karin M (2010) Immunity, inflammation, and cancer. *Cell* 140(6):883–899
56. Takeda K, Akira S (2005) Toll-like receptors in innate immunity. *Int Immunol* 17(1):1–14
57. Kovach MA, Standiford TJ (2011) Toll like receptors in diseases of the lung. *Int Immunopharmacol* 11(10):1399–1406
58. Medvedev AE (2013) Toll-like receptor polymorphisms, inflammatory and infectious diseases, allergies, and cancer. *J Interf Cytokine Res* 33(9):467–484

59. Misch EA, Hawn TR (2008) Toll-like receptor polymorphisms and susceptibility to human disease. *Clin Sci* 114(5):347–360
60. Netea MG, Wijmenga C, O'Neill LA (2012) Genetic variation in toll-like receptors and disease susceptibility. *Nat Immunol* 13(6):535–542
61. O'Neill LA (2003) Therapeutic targeting of toll-like receptors for inflammatory and infectious diseases. *Curr Opin Pharmacol* 3(4):396–403
62. Schwartz DA, Cook DN (2005) Polymorphisms of the toll-like receptors and human disease. *Clin Infect Dis* 7(29):S403–S407
63. Ermolaeva MA, Schumacher B (2014) Insights from the worm: the *C. elegans* model for innate immunity. *Semin Immunol* 26(4):303–309
64. Irazoqui J, Ausubel F (2010) 99th Dahlem conference on infection, inflammation, and chronic inflammatory disorders: *Caenorhabditis elegans* as a model to study tissues involved in host immunity and microbial pathogenesis. *Clin Exp Immunol* 160:48–57
65. Pukkila-Worley R, Ausubel FM (2012) Immune defense mechanisms in the *Caenorhabditis elegans* intestinal epithelium. *Curr Opin Immunol* 24(1):3–9
66. Mahajan-Miklos S, Tan MW, Rahme LG, Ausubel FM (1999) Molecular mechanisms of bacterial virulence elucidated using a *Pseudomonas aeruginosa*-*Caenorhabditis elegans* pathogenesis model. *Cell* 96(1):47–56
67. Tan MW, Mahajan-Miklos S, Ausubel FM (1999) Killing of *Caenorhabditis elegans* by *Pseudomonas aeruginosa* used to model mammalian bacterial pathogenesis. *Proc Natl Acad Sci U S A* 96(2):715–720
68. Tan MW, Rahme LG, Sternberg JA, Tompkins RG, Ausubel FM (1999) *Pseudomonas aeruginosa* killing of *Caenorhabditis elegans* used to identify *P. aeruginosa* virulence factors. *Proc Natl Acad Sci U S A* 96(5):2408–2413
69. Cohen LB, Troemel ER (2015) Microbial pathogenesis and host defense in the nematode *C. elegans*. *Curr Opin Microbiol* 23:94–101
70. Diogo J, Bratanich A (2014) The nematode *Caenorhabditis elegans* as a model to study viruses. *Arch Virol* 159(11):2843–2851
71. Arvanitis M, Glavis-Bloom J, Mylonakis E (2013) Invertebrate models of fungal infection. *Biochim Biophys Acta* 1832(9):1378–1383
72. Darby C (2005) Interactions with microbial pathogens. *WormBook: the online review of C elegans biology*, Pasadena (CA), pp 1–15
73. Dierking K, Yang W, Schulenburg H (2016) Antimicrobial effectors in the nematode *Caenorhabditis elegans*: an outgroup to the Arthropoda. *Philos Trans R Soc Lond Ser B Biol Sci* 371:20150299
74. Kim DH, Ewbank JJ (2015) Signaling in the innate immune response. *WormBook: the online review of C elegans biology*, Pasadena (CA), pp 1–51
75. Irazoqui JE, Troemel ER, Feinbaum RL, Luhachack LG, Cezairliyan BO, Ausubel FM (2010) Distinct pathogenesis and host responses during infection of *C. elegans* by *P. aeruginosa* and *S. aureus*. *PLoS Pathog* 6:e1000982
76. Visvikis O, Ihuegbu N, Labeled SA, Luhachack LG, Alves AM, Wollenberg AC, Stuart LM, Stormo GD, Irazoqui JE (2014) Innate host defense requires TFEB-mediated transcription of cytoprotective and antimicrobial genes. *Immunity* 40(6):896–909
77. Pastore N, Brady OA, Diab HI, Martina JA, Sun L, Huynh T, Lim JA, Zare H, Raben N, Ballabio A, Puertollano R (2016) TFEB and TFE3 cooperate in the regulation of the innate immune response in activated macrophages. *Autophagy* 12(8):1240–1258
78. Najibi M, Labeled SA, Visvikis O, Irazoqui JE (2016) An evolutionarily conserved PLC-PKD-TFEB pathway for host defense. *Cell Rep* 15(8):1728–1742
79. Alper S, McBride SJ, Lackford B, Freedman JH, Schwartz DA (2007) Specificity and complexity of the *C. elegans* innate immune response. *Mol Cell Biol* 27(15):5544–5553
80. Pulak R (2006) Techniques for analysis, sorting, and dispensing of *C. elegans* on the COPAS flow-sorting system. *Methods Mol Biol* 351:275–286
81. Alper S, Laws R, Lackford B, Boyd WA, Dunlap P, Freedman JH, Schwartz DA (2008) Identification of innate immunity genes and pathways using a comparative genomics approach. *Proc Natl Acad Sci U S A* 105(19):7016–7021
82. De Arras L, Laws R, Leach SM, Pontis K, Freedman JH, Schwartz DA, Alper S (2014) Comparative genomics RNAi screen identifies *Eftud2* as a novel regulator of innate immunity. *Genetics* 197(2):485–496
83. De Arras L, Seng A, Lackford B, Keikhae MR, Bowerman B, Freedman JH, Schwartz DA, Alper S (2013) An evolutionarily

- conserved innate immunity protein interaction network. *J Biol Chem* 288 (3):1967–1978
84. De Arras L, Yang IV, Lackford B, Riches DW, Prekeris R, Freedman JH, Schwartz DA, Alper S (2012) Spatiotemporal inhibition of innate immunity signaling by the Tbc1d23 RAB-GAP. *J Immunol* 188(6):2905–2913
 85. Editorial (2003) Whither RNAi? *Nat Cell Biol* 5(6):489–490
 86. De Arras L, Alper S (2013) Limiting of the innate immune response by SF3A-dependent control of MyD88 alternative mRNA splicing. *PLoS Genet* 9(10):e1003855
 87. O'Connor BP, Danhorn T, De Arras L, Flatley BR, Marcus RA, Farias-Hesson E, Leach SM, Alper S (2015) Regulation of toll-like receptor signaling by the SF3a mRNA splicing complex. *PLoS Genet* 11(2):e1004932
 88. Flegal KM, Carroll MD, Kit BK, Ogden CL (2012) Prevalence of obesity and trends in the distribution of body mass index among US adults, 1999–2010. *JAMA* 307(5):491–497
 89. Haslam DW, James WP (2005) Obesity. *Lancet* 366(9492):1197–1209
 90. Bleich S, Cutler D, Murray C, Adams A (2008) Why is the developed world obese? *Annu Rev Public Health* 29:273–295
 91. Ashrafi K (2007) Obesity and the regulation of fat metabolism. *WormBook: the online review of C elegans biology, Pasadena (CA)*, pp 1–20
 92. Jones KT, Ashrafi K (2009) *Caenorhabditis elegans* as an emerging model for studying the basic biology of obesity. *Dis Model Mech* 2(5–6):224–229
 93. Zheng J, Greenway FL (2012) *Caenorhabditis elegans* as a model for obesity research. *Int J Obes* 36(2):186–194
 94. McKay RM, McKay JP, Avery L, Graff JM (2003) *C elegans*: a model for exploring the genetics of fat storage. *Dev Cell* 4 (1):131–142
 95. Ashrafi K, Chang FY, Watts JL, Fraser AG, Kamath RS, Ahringer J, Ruvkun G (2003) Genome-wide RNAi analysis of *Caenorhabditis elegans* fat regulatory genes. *Nature* 421 (6920):268–272
 96. Liu Z, Li X, Ge Q, Ding M, Huang X (2014) A lipid droplet-associated GFP reporter-based screen identifies new fat storage regulators in *C. elegans*. *J Genet Genomics* 41(5):305–313
 97. Klass MR (1983) A method for the isolation of longevity mutants in the nematode *Caenorhabditis elegans* and initial results. *Mech Ageing Dev* 22(3–4):279–286
 98. Friedman DB, Johnson TE (1988) A mutation in the age-1 gene in *Caenorhabditis elegans* lengthens life and reduces hermaphrodite fertility. *Genetics* 118(1):75–86
 99. Friedman DB, Johnson TE (1988) Three mutants that extend both mean and maximum life span of the nematode, *Caenorhabditis elegans*, define the age-1 gene. *J Gerontol* 43(4):B102–B109
 100. Kenyon C, Chang J, Gensch E, Rudner A, Tabtiang R (1993) A *C. elegans* mutant that lives twice as long as wild type. *Nature* 366 (6454):461–464
 101. Riddle DL, Swanson MM, Albert PS (1981) Interacting genes in nematode dauer larva formation. *Nature* 290(5808):668–671
 102. Hu PJ (2007) Dauer. *WormBook: the online review of C elegans biology, Pasadena (CA)*, pp 1–19
 103. Albert PS, Riddle DL (1988) Mutants of *Caenorhabditis elegans* that form dauer-like larvae. *Dev Biol* 126(2):270–293
 104. Golden JW, Riddle DL (1984) A pheromone-induced developmental switch in *Caenorhabditis elegans*: temperature-sensitive mutants reveal a wild-type temperature-dependent process. *Proc Natl Acad Sci U S A* 81 (3):819–823
 105. Murphy CT, PJ H (2013) Insulin/insulin-like growth factor signaling in *C. elegans*. *WormBook: the online review of C elegans biology, Pasadena (CA)*, pp 1–43
 106. Altintas O, Park S, Lee SJ (2016) The role of insulin/IGF-1 signaling in the longevity of model invertebrates, *C. elegans* and *D. melanogaster*. *BMB Rep* 49(2):81–92
 107. Giannakou ME, Partridge L (2007) Role of insulin-like signalling in drosophila lifespan. *Trends Biochem Sci* 32(4):180–188
 108. Hwangbo DS, Gershman B, MP T, Palmer M, Tatar M (2004) *Drosophila* dFOXO controls lifespan and regulates insulin signalling in brain and fat body. *Nature* 429(6991):562–566
 109. Tatar M, Kopelman A, Epstein D, MP T, Yin CM, Garofalo RS (2001) A mutant drosophila insulin receptor homolog that extends lifespan and impairs neuroendocrine function. *Science* 292(5514):107–110
 110. Clancy DJ, Gems D, Harshman LG, Oldham S, Stocker H, Hafen E, Leivers SJ, Partridge L (2001) Extension of life-span by loss of CHICO, a drosophila insulin receptor substrate protein. *Science* 292 (5514):104–106
 111. Kappeler L, De Magalhaes Filho C, Dupont J, Leneuve P, Cervera P, Perin L, Loudes C, Blaise A, Klein R, Epelbaum J, Le Bouc Y, Holzenberger M (2008) Brain IGF-1 receptors control mammalian growth and lifespan

- through a neuroendocrine mechanism. *PLoS Biol* 6(10):e254
112. Holzenberger M, Dupont J, Ducos B, Leneuve P, Geloën A, Even PC, Cervera P, Le Bouc Y (2003) IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice. *Nature* 421(6919):182–187
 113. Brooks-Wilson AR (2013) Genetics of healthy aging and longevity. *Hum Genet* 132(12):1323–1338
 114. Chung WH, Dao RL, Chen LK, Hung SI (2010) The role of genetic variants in human longevity. *Ageing Res Rev* 9(Suppl 1):S67–S78
 115. Anselmi CV, Malovini A, Roncarati R, Novelli V, Villa F, Condorelli G, Bellazzi R, Puca AA (2009) Association of the FOXO3A locus with extreme longevity in a southern Italian centenarian study. *Rejuvenation Res* 12(2):95–104
 116. Daumer C, Flachsbart F, Caliebe A, Schreiber S, Nebel A, Krawczak M (2014) Adjustment for smoking does not alter the FOXO3A association with longevity. *Age* 36(2):911–921
 117. Flachsbart F, Caliebe A, Kleindorp R, Blanche H, von Eller-Eberstein H, Nikolaus S, Schreiber S, Nebel A (2009) Association of FOXO3A variation with human longevity confirmed in German centenarians. *Proc Natl Acad Sci U S A* 106(8):2700–2705
 118. Kuningas M, Magi R, Westendorp RG, Slagboom PE, Remm M, van Heemst D (2007) Haplotypes in the human Foxo1a and Foxo3a genes; impact on disease and mortality at old age. *Eur J Hum Genet* 15(3):294–301
 119. Li Y, Wang WJ, Cao H, Lu J, Wu C, Hu FY, Guo J, Zhao L, Yang F, Zhang YX, Li W, Zheng GY, Cui H, Chen X, Zhu Z, He H, Dong B, Mo X, Zeng Y, Tian XL (2009) Genetic association of FOXO1A and FOXO3A with longevity trait in Han Chinese populations. *Hum Mol Genet* 18(24):4897–4904
 120. Lunetta KL, D'Agostino RB Sr, Karasik D, Benjamin EJ, Guo CY, Govindaraju R, Kiel DP, Kelly-Hayes M, Massaro JM, Pencina MJ, Seshadri S, Murabito JM (2007) Genetic correlates of longevity and selected age-related phenotypes: a genome-wide association study in the Framingham study. *BMC Med Genet* 8(Suppl 1):S13
 121. Pawlikowska L, Hu D, Huntsman S, Sung A, Chu C, Chen J, Joyner AH, Schork NJ, Hsueh WC, Reiner AP, Psaty BM, Atzmon G, Barzilai N, Cummings SR, Browner WS, Kwok PY, Ziv E, Study of Osteoporotic F (2009) Association of common genetic variation in the insulin/IGF1 signaling pathway with human longevity. *Ageing Cell* 8(4):460–472
 122. Soerensen M, Dato S, Christensen K, McGue M, Stevnsner T, Bohr VA, Christiansen L (2010) Replication of an association of variation in the FOXO3A gene with human longevity using both case-control and longitudinal data. *Ageing Cell* 9(6):1010–1017
 123. Soerensen M, Nygaard M, Dato S, Stevnsner T, Bohr VA, Christensen K, Christiansen L (2015) Association study of FOXO3A SNPs and aging phenotypes in Danish oldest-old individuals. *Ageing Cell* 14(1):60–66
 124. Willcox BJ, Donlon TA, He Q, Chen R, Grove JS, Yano K, Masaki KH, Willcox DC, Rodriguez B, Curb JD (2008) FOXO3A genotype is strongly associated with human longevity. *Proc Natl Acad Sci U S A* 105(37):13987–13992
 125. Suh Y, Atzmon G, Cho MO, Hwang D, Liu B, Leahy DJ, Barzilai N, Cohen P (2008) Functionally significant insulin-like growth factor I receptor mutations in centenarians. *Proc Natl Acad Sci U S A* 105(9):3438–3442
 126. Tazearslan C, Huang J, Barzilai N, Suh Y (2011) Impaired IGF1R signaling in cells expressing longevity-associated human IGF1R alleles. *Ageing Cell* 10(3):551–554
 127. Curran SP, Ruvkun G (2007) Lifespan regulation by evolutionarily conserved genes essential for viability. *PLoS Genet* 3(4):e56
 128. Hamilton B, Dong Y, Shindo M, Liu W, Odell I, Ruvkun G, Lee SS (2005) A systematic RNAi screen for longevity genes in *C. elegans*. *Genes Dev* 19(13):1544–1555
 129. Hansen M, Hsu AL, Dillin A, Kenyon C (2005) New genes tied to endocrine, metabolic, and dietary regulation of lifespan from a *Caenorhabditis elegans* genomic RNAi screen. *PLoS Genet* 1(1):119–128
 130. Samuelson AV, Carr CE, Ruvkun G (2007) Gene activities that mediate increased life span of *C. elegans* insulin-like signaling mutants. *Genes Dev* 21(22):2976–2994
 131. Ni Z, Lee SS (2010) RNAi screens to identify components of gene networks that modulate aging in *Caenorhabditis elegans*. *Brief Funct Genomics* 9(1):53–64
 132. Yanos ME, Bennett CF, Kaerberlein M (2012) Genome-wide RNAi longevity screens in *Caenorhabditis elegans*. *Curr Genomics* 13(7):508–518

133. Longo VD, Antebi A, Bartke A, Barzilai N, Brown-Borg HM, Caruso C, Curiel TJ, de Cabo R, Franceschi C, Gems D, Ingram DK, Johnson TE, Kennedy BK, Kenyon C, Klein S, Kopchick JJ, Lepperdinger G, Madeo F, Mirisola MG, Mitchell JR, Passarino G, Rudolph KL, Sedivy JM, Shadel GS, Sinclair DA, Spindler SR, Suh Y, Vijg J, Vinciguerra M, Fontana L (2015) Interventions to slow aging in humans: are we ready? *Aging Cell* 14(4):497–510
134. Riera CE, Dillin A (2015) Can aging be “drugged”? *Nat Med* 21(12):1400–1405
135. Weber JJ, Sowa AS, Binder T, Hubener J (2014) From pathways to targets: understanding the mechanisms behind polyglutamine disease. *Biomed Res Int* 2014:701758
136. Zhao XN, Usdin K (2015) The repeat expansion diseases: the dark side of DNA repair. *DNA Repair* 32:96–105
137. Olejniczak M, Urbanek MO, Krzyzosiak WJ (2015) The role of the immune system in triplet repeat expansion diseases. *Mediat Inflamm* 2015:873860
138. Lee DY, McMurray CT (2014) Trinucleotide expansion in disease: why is there a length threshold? *Curr Opin Genet Dev* 26:131–140
139. Iyer RR, Pluciennik A, Napierala M, Wells RD (2015) DNA triplet repeat expansion and mismatch repair. *Annu Rev Biochem* 84:199–226
140. Morley JF, Brignull HR, Weyers JJ, Morimoto RI (2002) The threshold for polyglutamine-expansion protein aggregation and cellular toxicity is dynamic and influenced by aging in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A* 99(16):10417–10422
141. Brignull HR, Moore FE, Tang SJ, Morimoto RI (2006) Polyglutamine proteins at the pathogenic threshold display neuron-specific aggregation in a pan-neuronal *Caenorhabditis elegans* model. *J Neurosci* 26(29):7597–7606
142. Faber PW, Alter JR, MacDonald ME, Hart AC (1999) Polyglutamine-mediated dysfunction and apoptotic death of a *Caenorhabditis elegans* sensory neuron. *Proc Natl Acad Sci U S A* 96(1):179–184
143. Nollen EA, Garcia SM, van Haaften G, Kim S, Chavez A, Morimoto RI, Plasterk RH (2004) Genome-wide RNA interference screen identifies previously undescribed regulators of polyglutamine aggregation. *Proc Natl Acad Sci U S A* 101(17):6403–6408
144. Caldwell GA, Cao S, Sexton EG, Gelwix CC, Bevel JP, Caldwell KA (2003) Suppression of polyglutamine-induced protein aggregation in *Caenorhabditis elegans* by torsin proteins. *Hum Mol Genet* 12(3):307–319
145. Wang H, Lim PJ, Yin C, Rieckher M, Vogel BE, Monteiro MJ (2006) Suppression of polyglutamine-induced toxicity in cell and animal models of Huntington’s disease by ubiquilin. *Hum Mol Genet* 15(6):1025–1041
146. Querfurth HW, LaFerla FM (2010) Alzheimer’s disease. *N Engl J Med* 362(4):329–344
147. Burns A, Iliffe S (2009) Alzheimer’s disease. *BMJ* 338:b158
148. Dujardin S, Colin M, Buec L (2015) Invited review: animal models of tauopathies and their implications for research/translation into the clinic. *Neuropathol Appl Neurobiol* 41(1):59–80
149. Wentzell J, Kretzschmar D (2010) Alzheimer’s disease and tauopathy studies in flies and worms. *Neurobiol Dis* 40(1):21–28
150. Kraemer BC, Zhang B, Leverenz JB, Thomas JH, Trojanowski JQ, Schellenberg GD (2003) Neurodegeneration and defective neurotransmission in a *Caenorhabditis elegans* model of tauopathy. *Proc Natl Acad Sci U S A* 100(17):9980–9985
151. Kraemer BC, Burgess JK, Chen JH, Thomas JH, Schellenberg GD (2006) Molecular pathways that influence human tau-induced pathology in *Caenorhabditis elegans*. *Hum Mol Genet* 15(9):1483–1496
152. Hannan SB, Drager NM, Rasse TM, Voigt A, Jahn TR (2016) Cellular and molecular modifier pathways in tauopathies: the big picture from screening invertebrate models. *J Neurochem* 137(1):12–25
153. Guthrie CR, Schellenberg GD, Kraemer BC (2009) SUT-2 potentiates tau-induced neurotoxicity in *Caenorhabditis elegans*. *Hum Mol Genet* 18(10):1825–1838
154. Guthrie CR, Greenup L, Leverenz JB, Kraemer BC (2011) MSUT2 Is a determinant of susceptibility to tau neurotoxicity. *Hum Mol Genet* 20(10):1989–1999
155. Alexander AG, Marfil V, Li C (2014) Use of *Caenorhabditis elegans* as a model to study Alzheimer’s disease and other neurodegenerative diseases. *Front Genet* 5:279
156. Ewald CY, Li C (2010) Understanding the molecular basis of Alzheimer’s disease using a *Caenorhabditis elegans* model system. *Brain Struct Funct* 214(2–3):263–283
157. Hassan WM, Dostal V, Huemann BN, Yerg JE, Link CD (2015) Identifying Abeta-specific pathogenic mechanisms using a nematode model of Alzheimer’s disease. *Neurobiol Aging* 36(2):857–866

158. Hassan WM, Merin DA, Fonte V, Link CD (2009) AIP-1 ameliorates beta-amyloid peptide toxicity in a *Caenorhabditis elegans* Alzheimer's disease model. *Hum Mol Genet* 18 (15):2739–2747
159. Link CD (2006) *C. elegans* models of age-associated neurodegenerative diseases: lessons from transgenic worm models of Alzheimer's disease. *Exp Gerontol* 41 (10):1007–1013
160. Wu Y, Wu Z, Butko P, Christen Y, Lambert MP, Klein WL, Link CD, Luo Y (2006) Amyloid-beta-induced pathological behaviors are suppressed by Ginkgo Biloba extract EGb 761 and ginkgolides in transgenic *Caenorhabditis elegans*. *J Neurosci* 26 (50):13102–13113
161. Levy-Lahad E, Wasco W, Poorkaj P, Romano DM, Oshima J, Pettingell WH, CE Y, Jondro PD, Schmidt SD, Wang K et al (1995) Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science* 269 (5226):973–977
162. Rogaev EI, Sherrington R, Rogaeva EA, Levesque G, Ikeda M, Liang Y, Chi H, Lin C, Holman K, Tsuda T et al (1995) Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature* 376(6543):775–778
163. Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, Ikeda M, Chi H, Lin C, Li G, Holman K, Tsuda T et al (1995) Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* 375(6534):754–760
164. Levitan D, Doyle TG, Brousseau D, Lee MK, Thinakaran G, Slunt HH, Sisodia SS, Greenwald I (1996) Assessment of normal and mutant human presenilin function in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A* 93 (25):14940–14944
165. Levitan D, Greenwald I (1995) Facilitation of lin-12-mediated signalling by sel-12, a *Caenorhabditis elegans* S182 Alzheimer's disease gene. *Nature* 377(6547):351–354
166. Wong PC, Zheng H, Chen H, Becher MW, Sirinathsinghji DJ, Trumbauer ME, Chen HY, Price DL, Van der Ploeg LH, Sisodia SS (1997) Presenilin 1 is required for Notch1 and Dll1 expression in the paraxial mesoderm. *Nature* 387(6630):288–292
167. Francis R, McGrath G, Zhang J, Ruddy DA, Sym M, Apfeld J, Nicoll M, Maxwell M, Hai B, Ellis MC, Parks AL, Xu W, Li J, Gurney M, Myers RL, Himes CS, Hiebsch R, Ruble C, Nye JS, Curtis D (2002) Aph-1 and pen-2 are required for notch pathway signaling, gamma-secretase cleavage of betaAPP, and presenilin protein accumulation. *Dev Cell* 3(1):85–97
168. Goutte C, Tsunozaki M, Hale VA, Priess JR (2002) APH-1 is a multipass membrane protein essential for the notch signaling pathway in *Caenorhabditis elegans* embryos. *Proc Natl Acad Sci U S A* 99(2):775–779
169. Luo WJ, Wang H, Li H, Kim BS, Shah S, Lee HJ, Thinakaran G, Kim TW, Yu G, Xu H (2003) PEN-2 and APH-1 coordinately regulate proteolytic processing of presenilin 1. *J Biol Chem* 278(10):7850–7854
170. Samii A, Nutt JG, Ransom BR (2004) Parkinson's disease. *Lancet* 363(9423):1783–1793
171. Shulman JM, De Jager PL, Feany MB (2011) Parkinson's disease: genetics and pathogenesis. *Annu Rev Pathol* 6:193–222
172. Davie CA (2008) A review of Parkinson's disease. *Br Med Bull* 86:109–127
173. Atik A, Stewart T, Zhang J (2016) Alpha-synuclein as a biomarker for Parkinson's disease. *Brain Pathol* 26(3):410–418
174. Schulz-Schaeffer WJ (2010) The synaptic pathology of alpha-synuclein aggregation in dementia with Lewy bodies, Parkinson's disease and Parkinson's disease dementia. *Acta Neuropathol* 120(2):131–143
175. Chege PM, McColl G (2014) *Caenorhabditis elegans*: a model to investigate oxidative stress and metal dyshomeostasis in Parkinson's disease. *Front Aging Neurosci* 6:89
176. van Ham TJ, Thijssen KL, Breitling R, Hofstra RM, Plasterk RH, Nollen EA (2008) *C. elegans* model identifies genetic modifiers of alpha-synuclein inclusion formation during aging. *PLoS Genet* 4(3):e1000027
177. Hamamichi S, Rivas RN, Knight AL, Cao S, Caldwell KA, Caldwell GA (2008) Hypothesis-based RNAi screening identifies neuroprotective genes in a Parkinson's disease model. *Proc Natl Acad Sci U S A* 105 (2):728–733
178. Kuwahara T, Koyama A, Koyama S, Yoshina S, Ren CH, Kato T, Mitani S, Iwatsubo T (2008) A systematic RNAi screen reveals involvement of endocytic pathway in neuronal dysfunction in alpha-synuclein transgenic *C. elegans*. *Hum Mol Genet* 17 (19):2997–3009

Microbiome Sequencing Methods for Studying Human Diseases

Rebecca M. Davidson and L. Elaine Epperson

Abstract

Over the last decade, biologists have come to appreciate that the human body is inhabited by thousands of bacterial species in diverse communities unique to each body site. Moreover, due to high-throughput sequencing methods for microbial characterization in a culture-independent manner, it is becoming evident that the microbiome plays an important role in human health and disease. This chapter focuses on the most common form of bacterial microbiome profiling, targeted amplicon sequencing of the 16S ribosomal RNA (rRNA) subunit encoded by 16S rDNA. We discuss important features for designing and performing microbiome experiments on human specimens, including experimental design, sample collection, DNA preparation, and selection of the 16S rDNA sequencing target. We also provide details for designing fusion primers required for targeted amplicon sequencing and selecting the most appropriate high-throughput sequencing platform. We conclude with a review of the fundamental concepts of data analysis and interpretation for these kinds of experiments. Our goal is to provide the reader with the essential knowledge needed to undertake microbiome experiments for application to human disease research questions.

Key words Microbiome, 16S rRNA, 16S rDNA, Targeted amplicon sequencing, Bacteria

1 Introduction

The microbiome refers to the collection of microorganisms present within a community. While microorganisms include bacteria, fungi, protozoa, algae, and viruses, this chapter focuses on bacterial populations present in human specimens. Traditional approaches for studying bacterial ecology have relied on culture-dependent laboratory methods that are time-consuming, species-specific, low throughput, and underrepresentative of the diversity within a sample [1]. Current molecular and genomics methodologies, however, allow parallel profiling of nearly all bacteria in a sample with a single culture-independent experiment. As a result, modern microbiome research has blossomed over the last decade as researchers have described diverse and unique bacterial communities in a range of environmental samples [2], human specimens, and tissue types

[3]. The Human Microbiome Project (HMP) alone collected and profiled over 4700 specimens sampled from 15 to 18 body sites of 242 adults between 2008 and 2012 [4]. In the context of human disease, researchers are characterizing microbiome profiles of healthy and sick individuals, animal models of disease, and environments in which humans may acquire infectious pathogens [5].

There are two major high-throughput sequencing approaches for microbiome profiling. The first uses *targeted amplicon* sequencing, in which a conserved DNA target, such as the 16S ribosomal RNA subunit (rRNA) encoded by ribosomal DNA (rDNA) and present across all bacteria, is amplified by polymerase chain reaction (PCR) and sequenced. In this method, each sequence read corresponds to a copy of bacterial rDNA that is annotated and counted. The second approach is *metagenomic* sequencing, in which the total genomic DNA isolated from a sample is sequenced using a shotgun approach, thus nonselectively capturing all DNA sequences present in the specimen. This method captures the most genomic information as it is not restricted to a single taxonomic kingdom, but can be confounded by overly abundant eukaryotic DNA from the human host, and therefore may require significant sequencing depth to assay the breadth of bacterial diversity in a sample. In contrast, targeted 16S rDNA amplicon sequencing requires relatively low read depth to profile a broad range of 16S sequence variants in a sample. Because 16S rDNA amplicon sequencing is straightforward and has been widely applied in studies of microbiota related to human disease, we will focus exclusively on this approach in this chapter. Excellent reviews of metagenomic sequencing can be found elsewhere [6].

2 Materials and Methods

2.1 Selecting the DNA Sequencing Target and High-Throughput Sequencing Platform

Targeted amplicon sequencing of the 16S rDNA takes advantage of the alternating hypervariable and conserved regions that occur throughout the 1.5 kb gene. This structure is conserved across a range of prokaryotes, including Archaea and Bacteria. The experimental approach is to design universal primers within the conserved regions of the 16S rDNA that flank one or more variable regions containing the genetic variation used to classify each read into a taxonomic unit. More genetic differences in the variable region lead to better taxonomic separation of sequencing reads, which are useful for data interpretation.

The 16S rDNA includes nine hypervariable regions (V1–V9), each differing in length, genetic variability, and phylogenetic resolution for identifying a broad spectrum of bacterial taxa [7–9]. Over the years, many sets of universal primers have been designed to cover various regions of the 16S rDNA, and in the literature, they are named for their numerical position in the corresponding

Forward primer (515F): 5' GTG**Y**CAG**C**MGCCGCGGTAA 3'

Reverse primer (806R): 5' GGACTAC**N**VGGGT**W**TCTAAT 3'

Fig. 1 Example of universal primer sequences used for targeted 16S rDNA amplicon sequencing of the V4 region. Degenerate bases are highlighted in *red*

sequence of *Escherichia coli*. Regions V4, V5, and V6/7 have been shown to have the highest phylogenetic resolution, while regions V2 and V9 result in the poorest resolution [7, 10]. For universal primer sequences to amplify across a range of bacteria, they often contain degenerate bases. For example, in the primers that amplify the V4 region of the 16S rDNA (Fig. 1), there are 2/19 (11%) degenerate bases in the forward primer including **Y** (C or T) and **M** (A or C), and 3/20 (15%) degenerate bases in the reverse primer including **N** (A, C, G, or T), **V** (A, C, or G), and **W** (A or T).

The selection of primer sequences is generally based on two things: (1) taxonomic groups and the resolution desired for the sample type, and (2) compatibility of amplicon size with the chosen sequencing platform. For example, some common primer sets adapted for high-throughput sequencing include those designed for the Human Microbiome Project (HMP) that span the V1–V3 regions using primers 27F and 534R, and the V3–V5 regions using primers 357F and 926R (http://hmpdacc.org/doc/HMP_MDG_454_16S_Protocol.pdf). Phylogenetically, the V1–V3 region resolves many families of bacteria, but is less capable of distinguishing among Archaea. In contrast, the V3–V5 region provides high specificity across both prokaryotic kingdoms, but is not capable of resolving genera within the bacterial families Enterobacteriaceae and Pseudomonadaceae that include important human pathogens such as *Escherichia coli*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa*. The HMP primer sets yield amplicons greater than 500 bp and were designed in 2008 for use on the Roche-454 FLX Titanium sequencing platform [11]. In recent years, researchers have sacrificed the read length advantage of 454 sequencing for cost effectiveness and high read depth of alternative platforms such as MiSeq or HiSeq (Illumina) and the Ion Torrent Personal Genome Machine (Life Technologies).

Illumina sequencers provide single- or paired-end reads of 100 bp, 150 bp, or 300 bp, while the Ion Torrent has single-end 200 bp and 400 bp sequencing chemistries. Widely used primer sequences adapted for high-throughput Illumina sequencing were designed by the Earth Microbiome Project (EMP) consortium [2] and specifically target the V4 region [12]. Standard protocols for 16S amplicon sequencing using this method are available online at the EMP website (<http://www.earthmicrobiome.org/emp-standard-protocols/16s/>) and include primer sequences and PCR conditions. For the Ion Torrent PGM platform, primers in the V3

region (341F and 518R) and the V6 region (967F and 1046R) [13] are available, and Life Technologies recently released a 16S Metagenomics Kit that allows amplification and sequencing of seven hypervariable regions of the 16S rDNA in a single experiment. While this kit was evaluated to compare phylogenetic resolution of different amplicon regions, the potential advantage of combining sequence data for all hypervariable regions in a single analysis was not determined [10]. Theoretically, the extended sequence information that combines all of the hypervariable regions should be able to discern more taxa than any region alone.

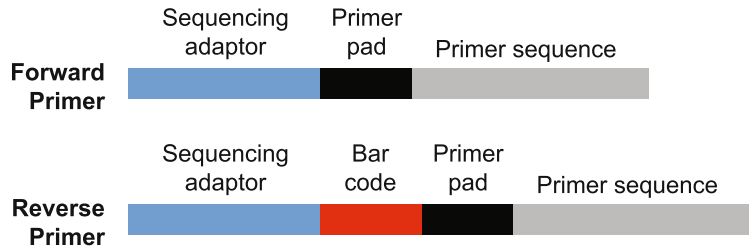
Library preparation for shotgun DNA sequencing typically comprises the following steps: (1) shearing DNA to the appropriate size, (2) repairing the ends of the sheared DNA, and (3) ligating sequencing adapters and barcodes onto DNA fragments. Sequencing adapters act as primers for the templating reactions prior to sequencing and are the same for each sample. Barcodes are unique sequence tags added to each DNA sample that allow computational separation of reads of pooled samples following sequencing and are used to increase throughput.

Amplicon sequencing is a relatively straightforward sequencing method because it does not require a number of the labor-intensive steps of the shotgun library preparation described above. Instead, the amplicon size is based on selection of universal primers and hypervariable region(s), and adapters and barcodes are incorporated during PCR using fusion primers. Fusion primer design is dependent on the sequencing platform, and some examples of fusion primer designs are shown in Fig. 2. For Illumina sequencing (Fig. 2a), both the forward and reverse primers contain sequencing adapters and primer pads, but only the reverse primer contains the 12 bp Golay barcode sequence [12]. The primer pad is a 10 bp sequence that prevents hairpin formation of the oligonucleotide [14]. For Ion Torrent sequencing (Fig. 2b), both the forward and reverse primers contain sequencing adapters, but it is the forward primer that contains the 10–12 bp barcode sequence, in addition to a 6 bp key sequence that is recognized by the Ion Torrent data processing software during base calling. In both cases, the barcode is located on only one primer. Therefore, the researcher will utilize multiple forward primers (for Ion Torrent sequencing, for example), each with a unique barcode tag for each sample, but will use the same reverse primer for all samples and PCR reactions.

2.2 Experimental Design

In designing microbiome experiments, it is important to realize that there is high variability among individual human samples [15, 16]. Previous work has shown that as little as 10% of taxa may be shared across a given population, also known as the “core microbiome,” suggesting that microbiome profiles are highly personalized. Therefore, cross-sectional studies may require a large number of individuals to distinguish relationships among

A. Fusion primer design for Illumina sequencing



B. Fusion primer design for Ion Torrent sequencing

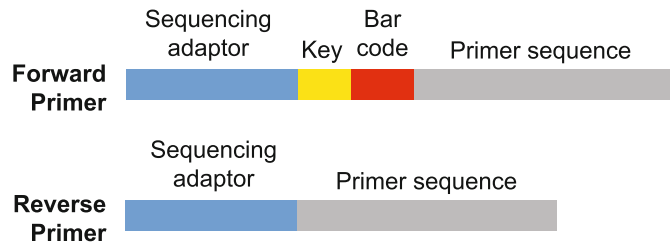


Fig. 2 Fusion primer design for targeted amplicon sequencing on the (a) Illumina MiSeq or HiSeq platforms or the (b) Ion Torrent Personal Genome Machine (PGM)

microbiota and clinical features or phenotypes. Otherwise, longitudinal study designs with fewer test subjects and extended time-points may be preferred. As with any biological dataset, careful attention should be paid to collecting accurate and complete metadata for each subject and sample so that robust statistical analyses may be performed. For microbiome studies in mouse models, additional factors such as age, genotype, and cohabitation must be considered, and comprehensive reviews on this topic are available [17, 18].

Microbiome sequencing is very sensitive to background contamination as only a few copies of DNA template are needed for PCR amplification [19, 20]. Thus, small quantities of bacterial contaminants present in ultrapure water, sterile saline, DNA isolation kits, or PCR reagents can show up in results from human-derived samples. Bacterial contamination from laboratory benches or sample handling can also contribute to biases in microbiome profiling results. This is especially a problem for low biomass samples, such as skin or airway specimens, as the DNA from the experimental sample is relatively less abundant than in other sample types (e.g., feces) relative to background contaminants [20]. This issue can be mitigated by using sterile technique while collecting and processing samples and by randomizing samples across different DNA extraction kits to minimize batch effect. In addition, the

inclusion of positive and negative controls, described in the section below, can allow for computational identification and correction for contaminating taxa in the final results.

2.3 Sample Preparation Considerations

To achieve thorough microbiome assessment, the essential starting point is a good preparation of DNA. There are several choices for DNA isolation, which include commercially available options such as the various column-based Qiagen, Zymo Research, and MO-BIO kits, or the traditional chemical and mechanical disruption methods, followed by phenol–chloroform extraction and ethanol precipitation. Perhaps the most important step in DNA isolation for microbiome sequencing is the lysis step, in which many types of bacteria with differing cell wall structures must be uniformly disrupted to release DNA for subsequent PCR amplification. This is generally accomplished with an enzymatic (lysozyme and/or proteinase K) or mechanical (bead-beating) step. The choice of cell lysis method will depend on the experimental question [21]; for example, recovery of DNA from acid-fast bacilli is improved by particular chemicals and specific bead sizes [22]. After lysis, the DNA should be recovered and contaminants removed to enable efficient PCR amplification. Column-based kits work well for this step. While some DNA may be lost to the column, phenolic compounds that impede PCR amplification can be avoided, which has the added benefit of not generating organic hazardous waste.

Microbiome experimentation is still in its infancy and universally accepted, standardized controls remain in a state of development. A number of negative controls are needed to address the background problem mentioned in the previous section. DNA should be prepared from water and any other solutions used throughout sample handling, and these samples need separate bar-coded libraries. The use of quantitative PCR (qPCR) is helpful to confirm that the DNA present in the samples is much more abundant than in the negative controls.

Positive controls are equally important for validation of microbiome results. The standard approach is to assemble a mock community of known DNAs in specified ratios comprising some of the target species of interest. Multiple mock communities with different known ratios are helpful in data interpretation because they reveal the ability of known sequences to amplify and be identified in an actual experiment, in addition to revealing any weaknesses in primer function. A mock community also allows the researcher to be confident that expected bacteria will be distinguished within the sample. Each mock community should have its own barcoded amplicon library that is prepared concomitantly with the sample libraries for any given experiment.

Ideally, PCR amplification of controls and samples should be executed in parallel, technical replicates with unique barcodes until reproducibility is established. The computational analysis of an

experiment is dependent on read counts obtained from the sequencing; for this reason, it is important to aim for equal loading of each library. The simplest way to normalize loading is to “clean” the amplified libraries using beads (e.g., Agencourt Ampure XP) to eliminate primers and other components of the PCR reaction that may interfere with sequencing. After size selection, the libraries are quantified using a fluorescence-based quantitation method (e.g., Qubit Fluorometric quantitation), and equal molar amounts of each sample combined to make the final library. The combined library is then requantified and the size distribution checked using gel electrophoresis or the Bioanalyzer instrument (Agilent). These measurements provide a good estimation of molarity necessary to proceed with sequencing of the combined libraries. Separate and combined libraries are stable at $-20\text{ }^{\circ}\text{C}$ for long-term storage or $4\text{ }^{\circ}\text{C}$ for several weeks to months.

2.4 Data Analysis Methods

The data output of 16S amplicon sequencing consists of one or more files in FASTQ format with hundreds to thousands of sequence reads. These reads are eventually converted into a matrix output in which the rows are 16S sequence variants, the columns are samples, and the cells comprise read counts for each 16S sequence variant by sample combination. There are several analysis steps that must occur to go from raw sequences to a matrix of annotated read counts for downstream analyses as shown in Fig. 3. Because of the relatively large size of amplicon sequencing datasets, analyses are usually performed on a high performance computing cluster with open-source, command line software packages such as QIIME [23] and Mothur [24]. Both packages offer modules and scripts for a range of analysis applications along with corresponding documentation that can be accessed online:

Steps for Microbiome Data Analyses

PRE-PROCESSING OF SEQUENCE READS

1. De-multiplex sequence reads by barcode
2. Trim and filter sequence reads for length and quality

OTU PICKING PIPELINE

3. Cluster sequence reads into OTUs based on 97% identity
4. Align representative OTU sequences for phylogenetic analyses
5. Assign taxonomy to OTU clusters by aligning to 16s rDNA database

DOWNSTREAM ANALYSES

6. Normalization – equal number of reads per sample (rarefaction)
7. Count matrices at phylum, class, order, family, genus and species levels
8. Community level analyses – i.e. PCoA, *alpha* and *beta* diversity
9. Statistical integration with clinical or disease phenotypes

Fig. 3 Steps for Microbiome Data Analyses

QIIME—<http://qiime.org/scripts/index.html>, Mothur—http://www.mothur.org/wiki/Main_Page.

Depending on the sequencing platform, the first step is to de-multiplex the sequence reads into separate read files for each sample. For Ion Torrent data, sequence reads are binned based on barcode sequence by the built-in software on the sequencing server, and data output are provided as separate FASTQ files for each sample. For the Illumina platform, reads may or may not be de-multiplexed on the sequencing server, and the researcher may need to separate the sequences by barcode using an open source module. Once the sequence reads have been separated by DNA sample, then quality filtering and trimming is used to remove low quality bases and short reads from the dataset, as these can be misannotated and may bias the overall results.

The next stage of analysis is read clustering in which reads are aligned to each other and grouped into clusters that share a minimum of 97% sequence identity based on the historical cutoff for bacterial species identification using the 16S rRNA gene [25]. These clusters, known as operational taxonomic units (OTUs), are computationally defined units, but are also thought to potentially represent unique biological entities. There are two flavors of OTU picking pipelines: (1) de novo OTU picking, in which all reads are included in the clustering analysis without initially referencing a database of known 16S rDNA sequences, and (2) open-reference OTU picking, in which reads are first aligned to a 16S rDNA sequence database to be categorized as known or novel, and then read clustering is performed separately for the two groups. The advantage of open reference compared to the de novo method is that some of the computing processes can be run in parallel, thereby reducing the overall computing time, while still retaining the novel OTUs acquired from the de novo method.

Once OTUs are picked, representative sequences are then chosen from each cluster to build alignment-based phylogenetic trees. This facilitates the generation of pairwise distance calculations between OTUs used in downstream analyses. Finally, each representative OTU sequence is compared to a curated rDNA sequence database, such as Greengenes [22], SILVA [23], or the RDP Classifier [24], and assigned a taxonomical annotation. It should be noted that only OTUs with significant matches to a database entry receive a complete taxonomical annotation, from phylum to species. Thus, some OTUs may only receive annotation at the phylum, class, order, or family level, with no genus or species classification.

The main output file from 16S microbiome analysis is the OTU count table, which is generated in the biological observation matrix (BIOM) format (<http://biom-format.org/>), a recognized standard of the Genomics Standards Consortium [26]. This standardized format is useful because it is compatible with a number of existing downstream analysis programs. After generating a BIOM table for

the sample set, the next step is to normalize read counts to equal numbers of reads per sample. A rarefaction analysis, or generation of rarefaction curves, is often useful to determine the minimum number of reads needed to detect the maximum species richness, or *alpha diversity*, within a sample, and can be performed using modules in QIIME or Mothur. This number, however, is more often defined or limited by the sample with the least number of reads in a given experiment. It should be noted that microbiome data can be analyzed with as few as 500 reads per sample to as many as 10,000 or more, although the ability to detect rare taxa will decrease with fewer reads. The normalized results are presented as relative abundance or percent of total reads for each taxonomic group observed in the community. This means that a microbiome profile generated with targeted amplicon sequencing does not represent the bacterial load or absolute abundance of bacteria in a sample, but rather the community composition in relative proportions. The addition of qPCR data of the 16S gene target can potentially be used to correct for absolute abundance, though this method has not been widely adopted.

Common downstream analyses of normalized microbiome data include comparing bacterial communities to each other and between sample subgroups. This is often done with measures of *beta diversity*, such as weighted and unweighted Unifrac or Bray Curtis dissimilarity metrics [27, 28]. These methods generate different forms of dissimilarity matrices among samples that can be visualized as dendrograms or in two-dimensional principle coordinates analysis (PCoA) plots. These analyses can be performed with QIIME or Mothur and visualized in programs such as phyloSeq [29].

According to data sharing practices for sequencing data, targeted amplicon sequences must be made publically available upon publication in a scientific journal. The Genomics Standards Consortium has developed a checklist of the minimum information about a marker gene sequence (MIMARKS) that is recommended before submission to a sequence database repository such as the National Center for Biotechnology Information (NCBI) [30]. This standardization provides researchers with enough information to confidently utilize publically available data for computational experiments or for comparisons with their own datasets.

3 Conclusions

Microbiome research has exploded over the last decade largely due to the relative ease and high-throughput nature of targeted 16S rDNA amplicon sequencing described in this chapter. Researchers with molecular capabilities can easily employ this PCR method in their laboratory and outsource the sequencing and primary analysis

steps to Genomics and Bioinformatics Core facilities. In many cases, a successful microbiome study is a multidisciplinary, collaborative effort between basic researchers, such as biologists or immunologists, and computational scientists, such as bioinformaticians and statisticians.

In the context of human diseases, much progress has been made in describing baseline features of the gut microbiome and how it is affected by factors such as age, sex, diet, and environmental exposures [4, 31]. The human gut microbiome can be categorized into three enterotypes based on the dominant genera present, namely *Bacteroides*, *Prevotella*, and *Ruminococcus* [32], and has been shown to be intimately connected to the mucosal immune system, as dysbiosis has been observed in patients with autoimmune disorders such as rheumatoid arthritis, multiple sclerosis, type 1 diabetes, and inflammatory bowel disease [31, 33]. Gut microbiota specifically interact with the immune system through cellular components and secreted metabolites, which influence the inflammation state of the intestinal mucosa [33]. Most research to date has focused on delineating pathogenic bacterial taxa from beneficial ones to inform potential therapeutic interventions, such as prebiotic and probiotic treatments. Much has also been learned by studying changes in the gut microbiome associated with successful treatment of ulcerative colitis [34] and intestinal infections of *Clostridium difficile* [35] using fecal microbiota transplantation.

Research on the lung microbiome suggests that it is remarkably stable, individualized, and relatively refractory to perturbations [36]. Methodological studies have evaluated the most informative and least invasive sample types to study [37]; airway samples include bronchoalveolar lavage (BAL), bronchial or nasal brushings, sputum, and biopsy. Disease states tend to correlate with greater bacterial load and enrichment or depletion of specific taxa. For example, the healthy lung is characterized by prevalence of *Prevotella* and *Veillonella*, both of which are reduced in the COPD lung [38]. Sputa from cystic fibrosis patient samples often contain an abundance of *Pseudomonas*, *Staphylococcus*, *Haemophilus*, and *Burkholderia*, and their relative composition and overall load vary in parallel with exacerbations and antibiotic treatment [39], patient age, stage of the disease [40], and CFTR genotype [41]. Asthmatic and airway allergic responses have been associated with microbial composition in the gastrointestinal tract, as demonstrated in germ-free mouse models [42]. Development of the host immune system depends on microbial composition of the environment, including the presence of pets in the home, for example [43].

The current challenges facing microbiome research are its relatively descriptive nature and the difficulty in establishing causal relationships between disease state, the patient's environment, and the bacterial taxonomic profiles in a given sample. Future studies will need to address the analytical challenges of

incorporating ecological, community-based data with clinical variables and outcomes to yield meaningful interpretations and preventative interventions, ideally using prospective study designs [44]. However, the integration of microbiome data with human genetic, metabolomics, and other large-scale data types will likely improve our understanding of the role of the microbiome in human health [45–48].

References

1. Bertelli C, Greub G (2013) Rapid bacterial genome sequencing: methods and applications in clinical microbiology. *Clin Microbiol Infect* 19(9):803–813. <https://doi.org/10.1111/1469-0691.12217>
2. Gilbert JA, Jansson JK, Knight R (2014) The earth microbiome project: successes and aspirations. *BMC Biol* 12:69. <https://doi.org/10.1186/s12915-014-0069-1>
3. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI (2007) The human microbiome project. *Nature* 449(7164):804–810. <https://doi.org/10.1038/nature06244>
4. Human Microbiome Project C (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486(7402):207–214. <https://doi.org/10.1038/nature11234>
5. Feazel LM, Baumgartner LK, Peterson KL, Frank DN, Harris JK, Pace NR (2009) Opportunistic pathogens enriched in showerhead biofilms. *Proc Natl Acad Sci U S A* 106(38):16393–16399. <https://doi.org/10.1073/pnas.0908446106>. 0908446106 [pii]
6. Ghurye JS, Cepeda-Espinoza V, Pop M (2016) Metagenomic assembly: overview, challenges and applications. *Yale J Biol Med* 89(3):353–362
7. Yang B, Wang Y, Qian PY (2016) Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics* 17:135. <https://doi.org/10.1186/s12859-016-0992-y>
8. Claesson MJ, Wang Q, O’Sullivan O, Greene-Diniz R, Cole JR, Ross RP, O’Toole PW (2010) Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Res* 38(22):e200. <https://doi.org/10.1093/nar/gkq873>
9. Chakravorty S, Helb D, Burday M, Connell N, Alland D (2007) A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods* 69(2):330–339. <https://doi.org/10.1016/j.mimet.2007.02.005>
10. Barb JJ, Oler AJ, Kim HS, Chalmers N, Wallen GR, Cashion A, Munson PJ, Ames NJ (2016) Development of an analysis pipeline characterizing multiple hypervariable regions of 16S rRNA using mock samples. *PLoS One* 11(2):e0148047. <https://doi.org/10.1371/journal.pone.0148047>
11. Human Microbiome Project C (2012) A framework for human microbiome research. *Nature* 486(7402):215–221. <https://doi.org/10.1038/nature11209>
12. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* 6(8):1621–1624. <https://doi.org/10.1038/ismej.2012.8>
13. Whiteley AS, Jenkins S, Waite I, Kresoje N, Payne H, Mullan B, Allcock R, O’Donnell A (2012) Microbial 16S rRNA ion tag and community metagenome sequencing using the ion torrent (PGM) platform. *J Microbiol Methods* 91(1):80–88. <https://doi.org/10.1016/j.mimet.2012.07.008>
14. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* 79(17):5112–5120. <https://doi.org/10.1128/AEM.01043-13>
15. Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, Knight R, Ley RE (2014) Conducting a microbiome study. *Cell* 158(2):250–262. <https://doi.org/10.1016/j.cell.2014.06.037>
16. Meisel JS, Hannigan GD, Tyldsley AS, SanMiguel AJ, Hodkinson BP, Zheng Q, Grice EA (2016) Skin microbiome surveys are strongly influenced by experimental design. *J Invest*

- Dermatol* 136(5):947–956. <https://doi.org/10.1016/j.jid.2016.01.016>
17. Laukens D, Brinkman BM, Raes J, De Vos M, Vandenebeele P (2016) Heterogeneity of the gut microbiome in mice: guidelines for optimizing experimental design. *FEMS Microbiol Rev* 40(1):117–132. <https://doi.org/10.1093/femsre/fuv036>
 18. Moore RJ, Stanley D (2016) Experimental design considerations in microbiota/inflammation studies. *Clin Transl Immunol* 5(7):e92. <https://doi.org/10.1038/cti.2016.41>
 19. Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R (2014) Tracking down the sources of experimental contamination in microbiome studies. *Genome Biol* 15(12):564. <https://doi.org/10.1186/s13059-014-0564-2>
 20. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 12:87. <https://doi.org/10.1186/s12915-014-0087-z>
 21. Yuan S, Cohen DB, Ravel J, Abdo Z, Forney LJ (2012) Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS One* 7(3):e33865. <https://doi.org/10.1371/journal.pone.0033865>
 22. Kaeser M, Ruf MT, Hauser J, Pluschke G (2010) Optimized DNA preparation from mycobacteria. *Cold Spring Harb Protoc* 2010(4):prot5408. <https://doi.org/10.1101/pdb.prot5408>
 23. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JL, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5):335–336. <https://doi.org/10.1038/nmeth.f.303>. <https://doi.org/10.1038/nmeth.f.303> [pii]
 24. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75(23):7537–7541. <https://doi.org/10.1128/AEM.01541-09>
 25. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311(5765):1283–1287. <https://doi.org/10.1126/science.1123061>
 26. McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F, Knight R, Caporaso JG (2012) The biological observation matrix (BIOM) format. *Gigascience* 1(1):7. <https://doi.org/10.1186/2047-217X-1-7>
 27. Wong RG, Wu JR, Gloor GB (2016) Expanding the unifracs toolbox. *PLoS One* 11(9):e0161196. <https://doi.org/10.1371/journal.pone.0161196>
 28. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R (2011) UniFrac: an effective distance metric for microbial community comparison. *ISME J* 5(2):169–172. <https://doi.org/10.1038/ismej.2010.133>
 29. McMurdie PJ, Holmes S (2013) Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8(4):e61217. <https://doi.org/10.1371/journal.pone.0061217>
 30. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, Morrison N, Rocca-Serra P, Sterk P, Arumugam M, Bailey M, Baumgartner L, Birren BW, Blaser MJ, Bonazzi V, Booth T, Bork P, Bushman FD, Buttigieg PL, Chain PS, Charlson E, Costello EK, Huot-Creasy H, Dawyndt P, DeSantis T, Fierer N, Fuhrman JA, Gallery RE, Gevers D, Gibbs RA, San Gil I, Gonzalez A, Gordon JL, Guralnick R, Hankeln W, Highlander S, Hugenholtz P, Jansson J, Kau AL, Kelley ST, Kennedy J, Knights D, Koren O, Kuczynski J, Kyrpides N, Larsen R, Lauber CL, Legg T, Ley RE, Lozupone CA, Ludwig W, Lyons D, Maguire E, Methe BA, Meyer F, Muegge B, Nakielny S, Nelson KE, Nemergut D, Neufeld JD, Newbold LK, Oliver AE, Pace NR, Palanisamy G, Peplies J, Petrosino J, Proctor L, Pruesse E, Quast C, Raes J, Ratnasingham S, Ravel J, Relman DA, Assunta-Sansone S, Schloss PD, Schriml L, Sinha R, Smith MI, Sodergren E, Spo A, Stombaugh J, Tiedje JM, Ward DV, Weinstock GM, Wendel D, White O, Whiteley A, Wilke A, Wortman JR, Yatsunenko T, Glockner FO (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIXS) specifications. *Nat Biotechnol* 29

- (5):415–420. <https://doi.org/10.1038/nbt.1823>
31. Rosser EC, Mauri C (2016) A clinical update on the significance of the gut microbiota in systemic autoimmunity. *J Autoimmun* 74:85–93. <https://doi.org/10.1016/j.jaut.2016.06.009>
 32. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos WM, Brunak S, Dore J, Meta HITC, Antolin M, Artiguenave F, Blottiere HM, Almeida M, Brechot C, Cara C, Chervaux C, Cultrone A, Delorme C, Denariac G, Dervyn R, Foerstner KU, Friss C, van de Guchte M, Guedon E, Haimet F, Huber W, van Hylckama-Vlieg J, Jamet A, Juste C, Kaci G, Knol J, Lakhdari O, Layec S, Le Roux K, Maguin E, Merieux A, Melo Minardi R, M'Rini C, Muller J, Oozeer R, Parkhill J, Renault P, Rescigno M, Sanchez N, Sunagawa S, Torrejon A, Turner K, Vandemeulebrouck G, Varela E, Winogradsky Y, Zeller G, Weissenbach J, Ehrlich SD, Bork P (2011) Enterotypes of the human gut microbiome. *Nature* 473(7346):174–180. <https://doi.org/10.1038/nature09944>
 33. Honda K, Littman DR (2012) The microbiome in infectious disease and inflammation. *Annu Rev Immunol* 30:759–795. <https://doi.org/10.1146/annurev-immunol-020711-074937>
 34. Ishikawa D, Sasaki T, Osada T, Kuwahara-Arai K, Haga K, Shibuya T, Hiramatsu K, Watanabe S (2016) Changes in intestinal microbiota following combination therapy with fecal microbial transplantation and antibiotics for ulcerative colitis. *Inflamm Bowel Dis* 23(1):116–125. <https://doi.org/10.1097/MIB.0000000000000975>
 35. Khoruts A, Dicksved J, Jansson JK, Sadowsky MJ (2010) Changes in the composition of the human fecal microbiome after bacteriotherapy for recurrent *Clostridium Difficile*-associated diarrhea. *J Clin Gastroenterol* 44(5):354–360. <https://doi.org/10.1097/MCG.0b013e3181c87e02>
 36. Carmody LA, Zhao J, Kalikin LM, LeBar W, Simon RH, Venkataraman A, Schmidt TM, Abdo Z, Schloss PD, LiPuma JJ (2015) The daily dynamics of cystic fibrosis airway microbiota during clinical stability and at exacerbation. *Microbiome* 3:12. <https://doi.org/10.1186/s40168-015-0074-9>
 37. Zemanick ET, Wagner BD, Robertson CE, Stevens MJ, Szefer SJ, Accurso FJ, Sagel SD, Harris JK (2014) Assessment of airway microbiota and inflammation in cystic fibrosis using multiple sampling methods. *Ann Am Thorac Soc* 12(2):221–229. <https://doi.org/10.1513/AnnalsATS.201407-310OC>
 38. Huffnagle GB (2016) Another piece in the “research mosaic” that describes the role of the lung microbiome in COPD. *Thorax* 71(9):777–778. <https://doi.org/10.1136/thoraxjnl-2015-207415>
 39. Stokell JR, Gharabeh RZ, Hamp TJ, Zapata MJ, Fodor AA, Steck TR (2015) Analysis of changes in diversity and abundance of the microbial community in a cystic fibrosis patient over a multiyear period. *J Clin Microbiol* 53(1):237–247. [https://doi.org/10.1128/JCM.02555-14.JCM.02555-14\[pii\]](https://doi.org/10.1128/JCM.02555-14.JCM.02555-14[pii])
 40. Coburn B, Wang PW, Diaz Caballero J, Clark ST, Brahma V, Donaldson S, Zhang Y, Surendra A, Gong Y, Elizabeth Tullis D, Yau YC, Waters VJ, Hwang DM, Guttman DS (2015) Lung microbiota across age and disease stage in cystic fibrosis. *Sci Rep* 5:10241. <https://doi.org/10.1038/srep10241>
 41. Cox MJ, Allgaier M, Taylor B, Baek MS, Huang YJ, Daly RA, Karaoz U, Andersen GL, Brown R, Fujimura KE, Wu B, Tran D, Koff J, Kleinhenz ME, Nielson D, Brodie EL, Lynch SV (2010) Airway microbiota and pathogen abundance in age-stratified cystic fibrosis patients. *PLoS One* 5(6):e11044. <https://doi.org/10.1371/journal.pone.0011044>
 42. Beck JM, Young VB, Huffnagle GB (2012) The microbiome of the lung. *Transl Res* 160(4):258–266. <https://doi.org/10.1016/j.trsl.2012.02.005>
 43. Noval Rivas M, Crother TR, Arditi M (2016) The microbiome in asthma. *Curr Opin Pediatr* 135(1):25–30. <https://doi.org/10.1097/MOP.0000000000000419>
 44. Hanson BM, Weinstock GM (2016) The importance of the microbiome in epidemiologic research. *Ann Epidemiol* 26(5):301–305. <https://doi.org/10.1016/j.annepidem.2016.03.008>
 45. Wu H, Tremaroli V, Backhed F (2015) Linking microbiota to human diseases: a systems biology perspective. *Trends Endocrinol Metab* 26(12):758–770. <https://doi.org/10.1016/j.tem.2015.09.011>

46. Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, Spector TD, Keinan A, Ley RE, Gevers D, Clark AG (2015) Host genetic variation impacts microbiome composition across human body sites. *Genome Biol* 16:191. <https://doi.org/10.1186/s13059-015-0759-1>
47. Dabrowska K, Witkiewicz W (2016) Correlations of host genetics and gut microbiome composition. *Front Microbiol* 7:1357. <https://doi.org/10.3389/fmicb.2016.01357>
48. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M, Van Treuren W, Knight R, Bell JT, Spector TD, Clark AG, Ley RE (2014) Human genetics shape the gut microbiome. *Cell* 159 (4):789–799. <https://doi.org/10.1016/j.cell.2014.09.053>

The Emerging Role of Long Noncoding RNAs in Human Disease

Johanna K. DiStefano

Abstract

Only a small fraction of the human genome corresponds to protein-coding genes. Historically, the vast majority of genomic sequence was dismissed as transcriptionally silent, but recent large-scale investigations have instead revealed a rich array of functionally significant elements, including non-protein-coding transcripts, within the noncoding regions of the human genome. Long noncoding RNAs (lncRNAs), a class of noncoding transcripts with lengths >200 nucleotides, are pervasively transcribed in the genome, and have been shown to bind DNA, RNA, and protein. lncRNAs exert effects through a variety of mechanisms that include guiding chromatin-modifying complexes to specific genomic loci, providing molecular scaffolds, modulating transcriptional programs, and regulating miRNA expression. An increasing number of experimental studies are providing evidence that lncRNAs mediate disease pathogenesis, thereby challenging the concept that protein-coding genes are the sole contributors to the development of human disease. This chapter highlights recent findings linking lncRNAs with human diseases of complex etiology, including hepatocellular carcinoma, Alzheimer's disease, and diabetes.

Key words Long noncoding RNAs, lncRNAs, Noncoding RNA, Hepatocellular carcinoma, Alzheimer's disease, Diabetes

1 lncRNAs: Challenging the Concept of Transcriptional Noise

Early large-scale studies estimated that approximately 5–10% of the human genome is transcribed, and only ~1% corresponds to protein-coding genes [1–3]. More recent findings from the Encyclopedia of DNA Elements (ENCODE) project reported that the number of GENCODE-annotated exons of protein-coding genes covers <3% of the entire genome, although greater than 80% of the human genome was found to engage in one or more biochemical RNA or chromatin-associated events [4]. Combined, these findings indicate that the vast majority of transcribed sequence is non-protein-coding. Noncoding elements comprise the majority of mammalian-conserved and recently adapted regions of the genome [5–8]. The overwhelming majority of genetic variants associated

with human traits or diseases have also been found to lie within intronic and intergenic regions [9], and are enriched within non-coding functional elements, particularly in regions located outside of those containing protein-coding genes [4]. The idea that non-coding regions of the human genome are transcriptionally silent and therefore, biologically irrelevant, has been updated with a better understanding of the deep complexity of our genetic code.

Some noncoding DNA is transcribed into functional noncoding RNA (ncRNA), a superclass of endogenous transcripts, which can be broadly classified based upon function into infrastructural and regulatory classes. Infrastructural (or housekeeping) ncRNA transcripts are involved in mRNA translation (ribosomal RNA and transfer RNA), as well as splicing and rRNA modification (small nuclear RNA and small nucleolar RNA, respectively), and are typically expressed in a constitutive manner. Regulatory ncRNAs mainly regulate gene expression and include long noncoding RNAs (lncRNAs) and short noncoding RNAs, such as microRNAs (miRNAs), Piwi-interacting RNAs (pi-RNAs), small interfering RNAs (siRNAs), promoter-associated RNAs (paRNAs), and enhancer RNAs (eRNAs). There are approximately 60,000 ncRNAs in the human genome, 68% of which are lncRNAs [10]. Because lncRNAs are typically expressed at much lower levels than mRNAs, these transcripts were considered to be transcriptional noise when first discovered in the early 1990s [11–13]. However, improvements in high-throughput sequencing technologies and computational methods have enabled the identification of a number of biologically relevant lncRNAs. Due to the emerging evidence supporting a substantial role for lncRNAs in biological processes underlying pathophysiology, this chapter focuses exclusively on these molecules within the context of human disease.

lncRNAs are a heterogeneous class of noncoding RNAs with transcript lengths >200 nucleotides [3]. Five major categories of lncRNAs have been defined relative to a spatial orientation with nearby protein-coding genes [14–17], and include (1) *intergenic* lncRNAs (lincRNAs), which are transcribed from regions at least >1 kb from protein-coding genes; (2) *bidirectional* lncRNAs, which are transcribed from regions within 1 kb of promoters in the opposite direction of the protein-coding transcript; (3) *intronic* lncRNAs, which are transcribed from introns of protein-coding genes [15, 16]; (4) *sense* lncRNAs, which are transcribed from the same strand of a protein-coding transcript and overlap with one or more exons of that transcript; and (5) *antisense* lncRNAs, which are transcribed from the opposite strand of a protein-coding gene and overlap with one or more exons of that transcript (Fig. 1). lncRNAs can exert *cis*- or *trans*-acting effects [18]. *Cis*-acting lncRNAs effects are limited to the chromosome from which they are transcribed and involve silencing or activation of gene expression. *Trans*-acting lncRNAs affect genes on chromosomes other

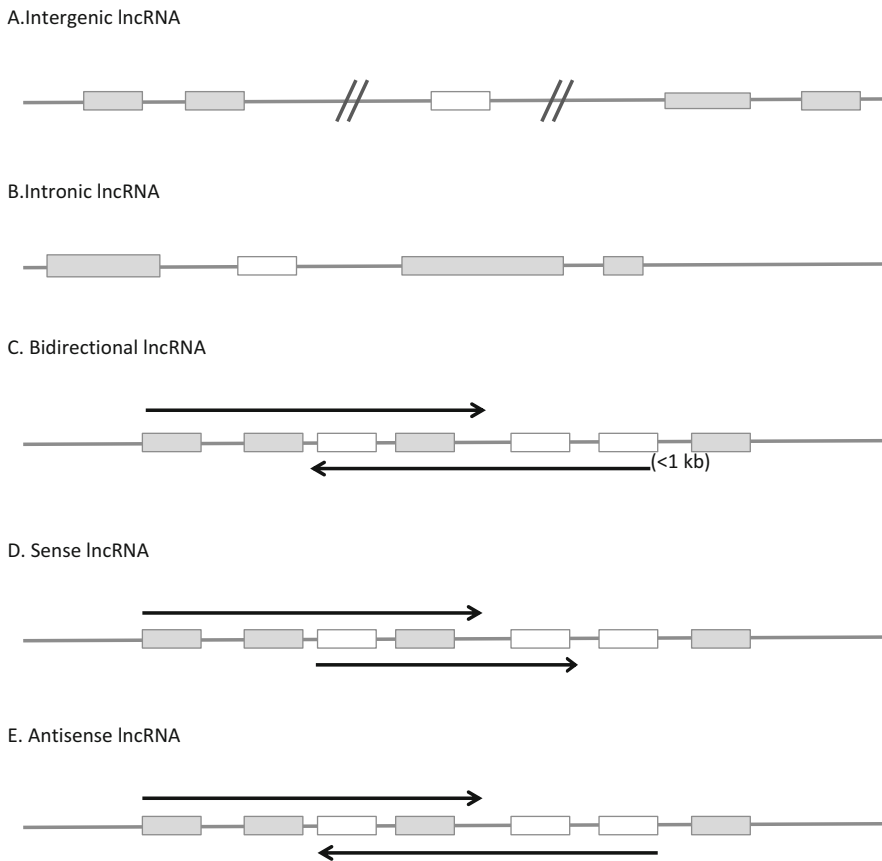


Fig. 1 The five main categories of lncRNAs. (a) Intergenic RNAs, (b) Intronic lncRNAs, (c) Bidirectional lncRNAs, (d) Sense lncRNAs, and (e) Antisense lncRNAs. Grey boxes represent protein-coding genes, with transcriptional direction depicted by arrows. Unfilled boxes depict lncRNAs. Diagonal lines in lincRNAs represent distances >1 kb. Adapted from [16](#)

than the one from which they are transcribed and regulate gene expression through recruitment of proteins to the target sites or sequestering of transcription factors away from targeted sites of transcription [19].

2 Characteristics and Functions of lncRNAs

Most lncRNAs are transcribed by RNA polymerase II, utilize the same consensus splicing signals as coding genes, and are posttranscriptionally modified at the 5' and 3' ends [20]. Despite these similarities with coding transcripts, lncRNAs tend to have shorter lengths and fewer exons, compared with mRNAs [14]. Conservation of lncRNAs across species is less than mRNAs [20, 21], although lncRNAs are likely to share similar functions [22, 23] and correlate with transposable elements, especially endogenous

retroviruses [24]. Compared with protein-coding genes, lncRNAs show stronger tissue-specific patterns of expression [25]. Many studies have reported that lncRNAs have very low expression levels [17]. However, a recent single-cell analysis of lncRNAs from the developing human cortex revealed abundant expression in individual cells compared to bulk tissue studies, suggesting that analysis of whole tissues may average gene expression signatures of many different cell types [26, 27], thereby muting actual expression patterns in individual cells.

lncRNAs have been shown to play a role in the regulation of gene expression, genomic imprinting, maintenance of pluripotency, nuclear organization and compartmentalization, and alternative splicing [16, 19, 28, 29]. In general, lncRNAs bind to DNA, RNA, and protein, and exert effects through these interactions. Mechanistically, lncRNAs can be classified into four major categories that include directing chromatin-modifying complexes to specific genomic loci, providing molecular scaffolds, modulating transcriptional programs, and regulating gene expression [16]. As shown in Fig. 2, these include (1) *decoy lncRNAs*, which regulate transcription by binding and sequestering a protein target, but do not exert any other effects; (2) *signal lncRNAs*, which regulate transcriptional activity or pathways in response to specific cues or stimuli; (3) *scaffold lncRNAs*, which act as platforms to host the formation of molecular complexes; and (4) *guide lncRNAs*, which bind protein and direct it to specific genomic loci. An excellent description of these mechanistic molecular functions, including lncRNA examples of each subclass, can be found elsewhere [19]. In addition to the functions listed above, lncRNAs can interact with miRNAs to exert effects on miRNA expression and activity [31]. These “sponges” limit the number of miRNA molecules available for binding with target genes [32], providing evidence that miRNA–lncRNA interactions represent an additional layer of transcriptional regulation within the cell [33].

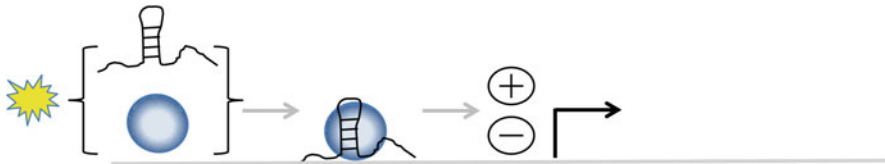
3 lncRNA–Disease Associations

Given the critical regulatory functions of lncRNAs in the cell, dysregulation of these molecules would be expected to contribute to pathophysiology. Indeed, according to the lncRNADisease database (<http://www.cuilab.cn/lncrnadisease>), a curated compilation of experimentally supported lncRNA–disease association data, there are more than 200 diseases associated with lncRNAs. A discussion of lncRNA involvement in three specific diseases: hepatocellular carcinoma, Alzheimer’s disease, and type 2 diabetes is presented in the following sections.

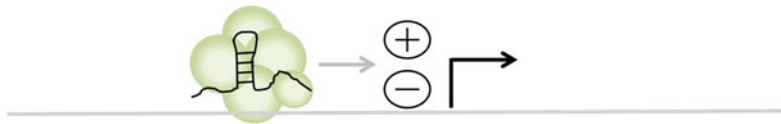
A. Decoy lncRNA



B. Signal lncRNA



C. Scaffold lncRNA



D. Guide lncRNA

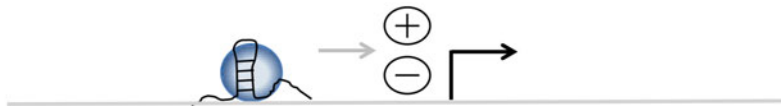


Fig. 2 Mechanistic categories of lncRNAs. (a) Decoy lncRNA, (b) Signal lncRNA, (c) Scaffold lncRNA, and (d) Guide lncRNA. *Hairpin loops* represent lncRNAs; *black arrows* represent transcriptional start sites; *blue and green circles* depict transcription factors and chromatin modifying complex, respectively; *yellow burst* indicates developmental or environmental signal. *Plus* and *minus* signs depict transcriptional activation or repression, respectively. Adapted from 30

3.1 lncRNAs and Hepatocellular Carcinoma: The Incredible HULC

Hepatocellular carcinoma is the most common form of liver cancer, and in the United States, the annual incidence of the disease is at least 6 per 100,000 [34]. The overall 5-year survival is less than 12%, making HCC the fastest rising cause of cancer-related death in the United States [35]. Risk factors for HCC development include viral infection, nonalcoholic fatty liver disease, alcohol overconsumption, aflatoxin, and genetic factors; the majority of these conditions contribute to the development of liver cirrhosis, which promotes HCC formation [35]. Over the past decade, a role for lncRNAs in the carcinogenesis, development, and prognosis of HCC has also emerged, and has developed into a highly active area of research. Although many lncRNAs have been identified [36–45], this chapter focuses on one specific lncRNA, HULC, because of the level of characterization achieved to date.

Comprehensive reviews of HCC-associated lncRNAs can be found elsewhere [46–49].

Panzitt et al. [50] performed a genome-wide search for novel transcripts associated with the molecular pathogenesis of HCC. Utilizing HCC-specific cDNA libraries and tissue samples from 46 HCCs, the authors identified a novel transcript that was upregulated 33-fold over a nonneoplastic pool of liver samples in 76% of tumors. This transcript was named HULC (highly upregulated in liver cancer). HULC expression was low in normal tissue and not significantly increased in other neoplastic tissues, suggesting that upregulation of this transcript was specific to HCC. In a study of 38 patients with HCC, HULC levels were also associated with clinical stage, intrahepatic metastases, HCC recurrence, and post-operative survival [51]. Knockdown of HULC expression in two hepatoma cells lines corresponded to dysregulation of many genes, including several with an established role in liver cancer; however, no significant sequence homology was observed between HULC and these potential target genes. These findings not only suggested that HULC has a general regulatory role as opposed to one specific gene target, but also indicated that the effect of HULC on potential downstream target genes is likely not based on direct RNA–RNA interactions.

Preliminary data showed that peripheral blood levels of HULC RNA were substantially higher in patients with HCC compared to individuals with no evidence of liver disease [50], a finding later corroborated in plasma samples of HCC patients [52]. Findings that HULC levels in blood mirror those in neoplasm indicate that this transcript may have utility as an improved and efficient noninvasive biomarker for the diagnosis and prognosis of HCC.

Transcription of HULC gives rise to a 482 bp, spliced, polyadenylated ncRNA that localizes to the cytoplasm and copurifies with ribosomes of carcinoma cells [50]. HULC was found to be evolutionarily conserved in primates, but neither mouse nor rat genomes appeared to have a HULC homolog [50]. Initial molecular characterization of HULC identified a cAMP response element binding (CREB) protein binding site in the proximal promoter region that was critical for HULC promoter activity in liver cancer [53]. In addition, HULC RNA was found to sequester miRNAs, including miR-372, which resulted in reduced expression of its target gene PRKACB [53].

Chronic infection with the hepatitis B virus (HBV) has been associated with the development of hepatocellular carcinoma for more than 3 decades [54]. Levels of HBV X protein (HBx), one of four proteins produced by HBV, are elevated in liver cells from patients with liver cancer [55], and HBx has been implicated in the development of HCC because it activates genes associated with cellular growth [56]. In an analysis of 33 clinical HCC specimens, expression levels of HBx and HULC were positively correlated

[57]. Knockdown of HBx expression resulted in decreased HULC expression in hepatoma cells, while overexpression of HBx led to a dose-dependent increase in HULC expression, suggesting that HBx regulates HULC expression in liver cells. Subsequent molecular characterization revealed a mechanism involving HBx-mediated activation via CREB binding of the HULC promoter [57]. HULC was found to downregulate p18, a tumor suppressor gene located in close proximity to HULC on chromosome 6p24.3, leading to proliferation of hepatoma cells. While HBx downregulates p18 in hepatoma cells, knockdown of HULC was able to rescue p18 expression levels, leading the authors to conclude that HBx-mediated upregulation of HULC promotes the proliferation of hepatoma cells through downregulation of p18 [57].

A recent study found that depletion of insulin growth factor 2 mRNA-binding protein 1 (IGF2BP1), but not IGF2BP2 or IGF2BP3, corresponded with an increased half-life and higher steady state transcript levels of HULC [58]. CNOT1 (C-C motif chemokine receptor 4 –NOT transcription complex subunit (1), a major component of the cytoplasmic RNA decay machinery, was identified as a novel interaction partner for IGF2BP1, and depletion of CNOT1 corresponded with increased half-life and expression of HULC. Thus, this study identified IGF2BP1 as an adaptor protein that recruits the CCR4-NOT complex, thereby initiating degradation of the HULC transcript [58].

Additional studies have served to shed light on the role of HULC in HCC development. For example, HULC was found to promote tumor angiogenesis by upregulating sphingosine kinase 1 [59], an enzyme that generates sphingosine phosphate, which promotes cell survival, proliferation, differentiation, and angiogenesis [60–62]. Through a series of experiments, the authors found that HULC increased expression of the transcription factor E2F1, which binds to the promoter of sphingosine kinase 1. HULC was found to sequester miR-107, which targets E2F1 for degradation through complementary base pairing in the 3' untranslated region, leading to upregulation of E2F1, increased activation of sphingosine kinase 1, and tumor angiogenesis. These findings provide new insight into mechanisms underlying tumor angiogenesis in HCC.

Likewise, HULC was recently found to enhance epithelial–mesenchymal transition, which contributes to tumor metastasis and recurrence in HCC via a signaling pathway involving zinc finger E-box binding homeobox 1 (ZEB1) and miR-200a-3p [51]. The authors demonstrated that HULC sequestered miR-200a-3p, leading to increased levels of ZEB1, which corresponded to stabilized epithelial–mesenchymal transition. These findings support a direct role for HULC in enhancing epithelial–mesenchymal transition, revealing potentially novel mechanisms by which the lncRNA mediates cancer pathophysiology in HCC.

3.2 *LncRNAs and Alzheimer's Disease*

Alzheimer's disease (AD) is a chronic neurodegenerative disease characterized by loss of neurons and synapses in the cerebral cortex, accumulation of amyloid plaques and neurofibrillary tangles, and progressive cognitive decline [63], although the disease is diagnosed primarily on the basis of extracellular plaque deposits of the β -amyloid peptide ($A\beta$) and the flame-shaped neurofibrillary tangles of the microtubule binding protein tau [64]. AD is considered to be the most common cause of dementia and affects about 46.8 million individuals worldwide. This figure is expected to double every 20 years, reaching 75 million by 2030 [65]. The average life expectancy post-diagnosis is 3–5 years [66, 67], and no therapies to stop or reverse the disease are available. Despite the public health importance of AD, the mechanisms underlying disease pathogenesis are poorly understood. However, a number of studies supporting a role for lncRNAs in the development and progression of AD have been accumulating in the literature over the past decade.

Brain cytoplasmic 200 RNA (BC200) is a translational regulator that targets eukaryotic initiation factor 4A, and plays roles in both regulating protein synthesis in postsynaptic dendritic microdomains and maintaining long-term synaptic plasticity [68]. An early investigation of BC200 identified a 70% reduction in levels of this ncRNA in a neocortical region known as Brodmann area 22 from AD patients compared to those of normal individuals [69]. In an independent analysis [70], BC200 were examined in Brodmann area 9, a prefrontal region in the superior frontal gyrus severely affected in patients with AD [71]. BC200 levels were significantly higher in AD tissue compared to normal tissue [70]. BC200 expression was reduced in aging normal brains, but significantly elevated in AD brains compared to age-matched normal brains. In the hippocampus, BC200 was also increased in AD brains relative to unaffected brains, but not in area 17 of the same samples. These results, combined with those from the earlier study [69], indicate that BC200 expression varies among different brain regions, and as such, may play a cell-specific role in the development of AD.

A noncoding antisense transcript of β -secretase 1 (BACE1) or BACE1-AS, was first identified as one of numerous sense–antisense transcript pairs conserved between humans and mice [72]. Subsequently, BACE1-AS levels were found to be elevated up to sixfold in various brain regions, including cerebellum, parietal lobe, hippocampus, superior frontal gyrus, and entorhinal cortex, of AD patients compared to age- and sex-matched brains [73]. Mechanistically, BACE1-AS and BACE1 were found to participate in a RNA duplex, which increased the stability of BACE1. BACE1-AS was also found to regulate expression of BACE1 mRNA and protein, and knockdown of BACE1-AS resulted in reduced levels of BACE1, $A\beta$ 1-40, and $A\beta$ 1-42, but not APP. Upon exposure to various cell stressors including amyloid- β 1-42 ($A\beta$ 1-42),

expression of *BACE1-AS* increased, leading to enhanced *BACE1* mRNA stability and generating additional A β 1-42 through a post-transcriptional feed-forward mechanism. These data show that *BACE1* mRNA expression is under the control of a regulatory noncoding RNA that may drive Alzheimer's disease-associated pathophysiology. The authors further showed that *BACE1-AS* prevents miRNA-induced translational repression and mRNA decay of *BACE1* mRNA by "masking" a binding site for miR-485-5p. *BACE1-AS* and miR-485-5p ncRNAs were found to compete with each other for binding to the sixth exonic region of *BACE1* mRNA [74]. Opposing regulatory effects of *BACE1-AS* and miR-485-5p on *BACE1* protein expression were also observed, suggesting the presence of a ncRNA regulatory network that controls *BACE1* expression, which, when altered, may be implicated in AD pathophysiology. Kang et al. [75] reported that the RNA-binding protein HuD, implicated in learning and memory, stabilized *BACE1-AS*, thereby enhancing *BACE1* expression. Levels of APP, *BACE1*, *BACE1AS*, and A β were elevated not only in brains of HuD-overexpressing mice, but also in the superior temporal gyrus of AD patients compared to age-matched control tissue.

In addition to BC200 and *BACE1-AS*, other lncRNAs have been associated with AD, including 17A, NDM29, and 51A [76–78]. LncRNA 17A, which is transcribed in an antisense orientation from the third intron of the G-protein-coupled receptor 51 (*GPR51*) gene was shown to regulate *GPR51* pre-mRNA processing and produce an alternative splicing isoform B of the GABA B receptor that abolishes GABA B2 intracellular signaling [78]. Levels of 17a were elevated in brain tissue from patients with AD compared to normal brain, and 17a expression in neuroblastoma cells corresponded with increased secretion of A β secretion and the A β \times -42/ ζ β \times -40 peptide ratio, a biomarker for AD. Synthesis of 17a was also found to increase in response to inflammatory stimuli. Likewise, an investigation of the neuroblastoma differentiation marker (*NDM29*) gene, which fosters differentiation of neuroblastoma cells to a nonmalignant phenotype, found that *NDM29*-dependent cell maturation corresponded with increased synthesis of amyloid precursor protein, resulting in enhanced A β secretion and elevation of the A β \times -42/A β \times -40 ratio [77]. *NDM29* expression was increased in cerebral tissues of AD patients and in response to inflammatory stimuli, leading to increased A β formation. LncRNAs 51A, which maps in an antisense orientation to intron 1 of the sortilin-related receptor 1 (*SORL1*), a risk gene for late-onset AD, was also found to be upregulated in cerebral cortices from AD patients [76]. Expression of 51A corresponds with an alternatively spliced *SORL1* isoform, which is associated with impaired processing of amyloid precursor protein (APP), leading to increased A β formation.

Several recent analyses of lncRNAs expression profiles in AD have been reported, providing additional insights in the etiology and pathophysiology of the disease. For example, microarray-based expression profiling of lncRNAs in a triple transgenic model of AD revealed 205 dysregulated lncRNAs in comparisons of affected and control mice [79]. Out of these dysregulated lncRNAs, 27 were located next to protein-coding genes that were also differentially expressed between AD and control animals, and many of the lncRNA-mRNA pairs were dysregulated in the same direction. A similar study applied microarray analysis to examine hippocampal lncRNAs expression in a rat model of AD, finding a total of 315 dysregulated lncRNAs [80]. In humans, microarray analysis of postmortem tissue samples from AD patients and age-matched controls revealed 108 differentially expressed lncRNAs [81]. An analysis of lncRNAs in various tissues indicated that most down-regulated lncRNAs in AD are highly expressed in the brain but not in other tissues. Gene set enrichment analysis identified a down-regulated lncRNA, n341006, associated with the protein ubiquitination pathway, and a significantly upregulated lncRNA, n336934, linked to cholesterol homeostasis. lncRNA expression signatures could predict tissue types with equivalent accuracy as protein-coding genes, but the number required for optimal prediction was less compared to mRNA signatures. Using RNA-sequencing of hippocampus samples from patients with late onset AD and age-matched controls, Magistri et al. [82] identified several annotated and nonannotated lncRNAs differentially expressed in brain tissues, three of which were activity-dependent regulated, and one induced by A β [1–29, 31–43] exposure of human neural cells. Despite the significance of findings in mice, rats, and humans, there is little overlap among studies, and further research is required to fully elucidate the detailed molecular mechanisms underlying the action of significantly dysregulated lncRNAs.

3.3 *LncRNAs and T2D Pathogenesis*

The role of ncRNAs in the pathogenesis of T2D has only recently become recognized, yet a growing list of lncRNAs involved in glucose homeostasis is emerging, as recently reviewed by Sun and Wong [83]. Here, we will briefly discuss two lncRNAs, H19 and MEG3, for which a substantial amount of experimental evidence has been reported.

H19 encodes a maternally expressed lncRNA [84] that plays a role in cell proliferation [85], regulation of gene expression, and development of some cancers [86]. H19 is located on chromosome 11p15.5, approximately 100 kb distal of insulin-like growth factor 2 (IGF2), and together H19 and IGF2 are transcribed from a conserved imprinted gene cluster [85]. Both H19 and IGF2 are abundantly expressed during fetal development, then downregulated in most tissues following birth, with the exceptions of skeletal muscle and heart. In mice, offspring with maternal deletion of H19

are significantly heavier than those inheriting a paternal deletion, although these findings were attributed to a gain-of-function of IGF2, which is paternally imprinted, rather than H19 loss-of-function [87]. Final trimester maternal glucose concentrations were significantly higher in mothers carrying pups with targeted disruption of H19 compared to wild-type animals [88]. Genetic variation in H19 was not associated with significant changes in maternal glucose tolerance in humans during the final trimester of pregnancy, although maternally transmitted H19 alleles were associated with increased birth weight, increased head circumference, and increased sum of skinfold thicknesses in offspring [89].

H19 levels are approximately five times lower in skeletal muscle of patients with T2D compared to healthy individuals, corresponding to increased bioavailability of the miRNA let-7 [90]. Under normal conditions, H19 sequesters let-7 [91], preventing it from binding to target genes, insulin receptor (INSR) and lipoprotein lipase (LPL) [91]. However, under diabetic conditions, in which decreased H19 leads to enhanced let-7 levels, expression of INSR and LPL is inhibited, leading to dysregulated glucose metabolism in skeletal muscle. Hyperinsulinemia was found to downregulate H19 expression through a pathway involving PI3K/AK-dependent phosphorylation of the miRNA-processing factor KSRP, which promotes let-7 biogenesis and subsequent H19 destabilization. Thus, these findings identified a double-negative feedback loop between H19 and let-7 for regulating glucose homeostasis in skeletal muscle.

In addition to H19, MEG3 (maternally expressed 3 gene) has emerged as an important player in glucose homeostasis. Like H19, MEG3 is a maternally expressed imprinted lncRNA [92], with established roles in cell proliferation [93–95]. MEG3 expression was upregulated in *ob/ob* mice, a model for T2D, and mice fed a high-fat diet, consisting of palmitate, oleate, or linoleate [96]. Overexpression of MEG3 corresponded to increased hepatic gluconeogenesis and suppressed insulin-stimulated glycogen synthesis in primary hepatocytes. In addition, levels of FOXO1, G6PC, and PEPCK increased in response to MEG3 overexpression, while levels of palmitate-induced FOXO1, G6PC, and PEPCK were reversed with MEG3 downregulation. The authors found that MEG3 knockdown could also reverse triglyceride upregulation, impaired glucose tolerance, and downregulation of glycogen content in high-fat diet-fed or *ob/ob* mice.

In addition to its role in mediating hepatic insulin resistance, MEG3 affects insulin synthesis and secretion in pancreatic β -cells [97]. MEG3 expression was significantly higher in islets compared to exocrine glands of *Balb/c* mice, but islet expression was reduced in NOD mice, a model of type 1 diabetes, and *db/db* mice, a model of T2D. In isolated mouse islets and a pancreatic β -cell line, MEG3 expression was regulated by glucose. Knockdown of MEG3 in vitro

led to impairment of insulin synthesis and secretion, and increased the rate of β -cell apoptosis, while in vivo knockdown resulted in impaired glucose tolerance and decreased insulin secretion, corresponding to reduced levels of insulin-positive cells. MEG3 knockdown also corresponded with decreased levels of the transcription factors, pancreatic and duodenal homeobox 1 (Pdx1) and v-maf musculoaponeurotic fibrosarcoma oncogene family, protein A (Maf). These results suggest that MEG3 may affect the development of diabetes via effects on β -cell maintenance and apoptosis.

Two recent studies examined global changes in lncRNAs expression relative to T2D. In a study of human pancreatic islet and β -cells, over 1000 lncRNAs were found to be islet-specific, compared to only 9.4% of RefSeq annotated genes [98]. Further, more than 19% of the transcribed genome in islets mapped outside of annotated protein-coding genes. The majority of lncRNAs identified were either silent or expressed at very low levels in pancreatic progenitors, but active in adult islets, indicating roles in pancreatic endocrine differentiation. Likewise, six lncRNAs expressed at very low or undetectable levels throughout in vitro differentiation became activated only during the in vivo maturation step [98]. In an investigation of 55 T2D susceptibility loci, nine contained islet lncRNAs within 150 kb of the reported lead marker, including six which have been linked directly to β -cell dysfunction [99–103]. Suppressed expression of HI-LNC25, a candidate lncRNA, in a human β -cell line, led to reduction in levels of GLIS3, an islet transcription factor mutated in monogenic diabetes.

In pancreatic islets from 89 donors with varying degrees of glucose tolerance, nearly 500 RefSeq islet lincRNAs were identified, 54 of which were associated with gene expression (eQTL) and exon use [104]. Seventeen lincRNAs were associated with HbA1c levels, including HI-LNC901 (i.e., LOC283177), which also had an eQTL (rs73036390) and whose expression was directly correlated with insulin exocytosis. HI-LNC901 was coexpressed with MAP-kinase activating death domain (MADD), synaptotagmin 11 (SYT11), and paired box 6 (PAX6), all of which have been implicated in islet function.

Other lncRNAs have been found to harbor genetic variants associated with T2D, including the *ANRIL* locus [105]. This lncRNA maps to the *INK4* locus and is required for the silencing of the p15^{INK4B} tumor suppressor gene [106]. Variants in ANRIL that disrupt its expression or function may affect compensatory increases in pancreatic β -cell mass in response to increasing demands for insulin in the pre-diabetes state [107].

4 Conclusions

The discovery of dysregulated lncRNAs contributes a new layer of complexity to the molecular architecture of human disease. However, there are still many gaps in our current understanding of lncRNA function, and further study of these molecules is expected to yield deeper insights into mechanisms underlying the pathogenesis of many human diseases, development of new RNA-based targets for the prevention and treatment of disease, and improved methods for early detection of pathology.

References

- Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science* 306:2242–2246
- Cheng P, Dolinsky V, Hatch GM (1996) The acylation of lysophosphatidylglycerol in rat heart: evidence for both in vitro and in vivo activities. *Biochim Biophys Acta* 1302:61–68
- Kapranov P, Cheng J, Dike S, Nix DA, Dutttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammanna H, Gingeras TR (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316:1484–1488
- Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74
- Fraser HB (2013) Gene expression drives local adaptation in humans. *Genome Res* 23:1089–1096
- Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, Cabili M, Adegbola RA, Bamezai RN, Hill AV, Vannberg FO, Rinn JL, Genomes P, Lander ES, Schaffner SF, Sabeti PC (2013) Identifying recent adaptations in large-scale genomic data. *Cell* 152:703–713
- Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alföldi J, Beal K, Chang J, Clawson H, Cuff J, Di Palma F, Fitzgerald S, Flicek P, Guttman M, Hubisz MJ, Jaffe DB, Jungreis I, Kent WJ, Kostka D, Lara M, Martins AL, Masingham T, Moltke I, Raney BJ, Rasmussen MD, Robinson J, Stark A, Vilella AJ, Wen J, Xie X, Zody MC, P. Broad Institute Sequencing, T. Whole Genome Assembly, Baldwin J, Bloom T, Chin CW, Heiman D, Nicol R, Nusbaum C, Young S, Wilkinson J, Worley KC, Kovar CL, Muzny DM, Gibbs RA, T. Baylor College of Medicine Human Genome Sequencing Center Sequencing, Cree A, Dihn HH, Fowler G, Jhangiani S, Joshi V, Lee S, Lewis LR, Nazareth LV, Okwuonu G, Santibanez J, Warren WC, Mardis ER, Weinstock GM, Wilson RK, U. Genome Institute at Washington, Delehaunty K, Dooling D, Fronik C, Fulton L, Fulton B, Graves T, Minx P, Sodergren E, Birney E, Margulies EH, Herrero J, Green ED, Haussler D, Siepel A, Goldman N, Pollard KS, Pedersen JS, Lander ES, Kellis M (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478:476–482
- Mouse Genome Sequencing C, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyraas E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA,

- Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korfi I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562
9. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 106:9362–9367
 10. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, Poliakov A, Cao X, Dhanasekaran SM, Wu YM, Robinson DR, Beer DG, Feng FY, Iyer HK, Chinnaiyan AM (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* 47:199–208
 11. Borsani G, Tonlorenzi R, Simmler MC, Dandolo L, Arnaud D, Capra V, Grompe M, Pizzuti A, Muzny D, Lawrence C, Willard HF, Avner P, Ballabio A (1991) Characterization of a murine gene expressed from the inactive X chromosome. *Nature* 351:325–329
 12. Brannan CI, Dees EC, Ingram RS, Tilghman SM (1990) The product of the H19 gene may function as an RNA. *Mol Cell Biol* 10:28–36
 13. Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S, Rastan S (1992) The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* 71:515–526
 14. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigo R (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 22:1775–1789
 15. Louro R, Smirnova AS, Verjovski-Almeida S (2009) Long intronic noncoding RNA transcription: expression noise or expression choice? *Genomics* 93:291–298
 16. Moran VA, Perera RJ, Khalil AM (2012) Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. *Nucleic Acids Res* 40:6391–6400
 17. Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. *Cell* 136:629–641
 18. Kornienko AE, Guenzl PM, Barlow DP, Pauller FM (2013) Gene regulation by the act of long non-coding RNA transcription. *BMC Biol* 11:59
 19. Wang KC, Chang HY (2011) Molecular mechanisms of long noncoding RNAs. *Mol Cell* 43:904–914
 20. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458:223–227
 21. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties

- and specific subclasses. *Genes Dev* 25:1915–1927
22. Diederichs S (2014) The four dimensions of noncoding RNA conservation. *Trends Genet* 30:121–123
 23. Gutschner T, Hammerle M, Eissmann M, Hsu J, Kim Y, Hung G, Revenko A, Arun G, Stenstrup M, Gross M, Zornig M, MacLeod AR, Spector DL, Diederichs S (2013) The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res* 73:1180–1189
 24. Johnson R, Guigo R (2014) The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA* 20:959–976
 25. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* 9:e1003470
 26. Liu SJ, Nowakowski TJ, Pollen AA, Lui JH, Horlbeck MA, Attenello FJ, He D, Weissman JS, Kriegstein AR, Diaz AA, Lim DA (2016) Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol* 17:67
 27. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, Trombetta JJ, Gennert D, Gnirke A, Goren A, Hacohen N, Levin JZ, Park H, Regev A (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498:236–240
 28. Nagano T, Fraser P (2011) No-nonsense functions for long noncoding RNAs. *Cell* 145:178–181
 29. Yan B, Wang Z (2012) Long noncoding RNA: its physiological and pathological roles. *DNA Cell Biol* 31(Suppl 1):S34–S41
 30. Lorenzen JM, Thum T (2016) Long noncoding RNAs in kidney and cardiovascular diseases. *Nat Rev Nephrol* 12:360–373
 31. Jalali S, Bhartiya D, Lalwani MK, Sivasubbu S, Scaria V (2013) Systematic transcriptome wide analysis of lncRNA-miRNA interactions. *PLoS One* 8:e53823
 32. Khorkova O, Hsiao J, Wahlestedt C (2015) Basic biology and therapeutic implications of lncRNA. *Adv Drug Deliv Rev* 87:15–24
 33. Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J (2013) Natural RNA circles function as efficient microRNA sponges. *Nature* 495:384–388
 34. El-Serag HB, Kanwal F (2014) Epidemiology of hepatocellular carcinoma in the United States: where are we? Where do we go? *Hepatology* 60:1767–1775
 35. Mittal S, El-Serag HB (2013) Epidemiology of hepatocellular carcinoma: consider the population. *J Clin Gastroenterol* 47(Suppl): S2–S6
 36. Jia M, Jiang L, Wang YD, Huang JZ, Yu M, Xue HZ (2016) lncRNA-p21 inhibits invasion and metastasis of hepatocellular carcinoma through notch signaling-induced epithelial-mesenchymal transition. *Hepatol Res* 46(11):1137–1144
 37. Peng W, Fan H (2016) Long noncoding RNA CCHE1 indicates a poor prognosis of hepatocellular carcinoma and promotes carcinogenesis via activation of the ERK/MAPK pathway. *Biomed Pharmacother* 83:450–455
 38. Sui CJ, Zhou YM, Shen WF, Dai BH, Lu JJ, Zhang MF, Yang JM (2016) Long noncoding RNA GIHCG promotes hepatocellular carcinoma progression through epigenetically regulating miR-200b/a/429. *J Mol Med (Berl)* 94(11):1281–1296
 39. Wang T, Ma S, Qi X, Tang X, Cui D, Wang Z, Chi J, Li P, Zhai B (2016) Long noncoding RNA ZNF1-AS1 suppresses growth of hepatocellular carcinoma cells by regulating the methylation of miR-9. *Onco Targets Ther* 9:5005–5014
 40. Xiong D, Sheng Y, Ding S, Chen J, Tan X, Zeng T, Qin D, Zhu L, Huang A, Tang H (2016) LINC00052 regulates the expression of NTRK3 by miR-128 and miR-485-3p to strengthen HCC cells invasion and migration. *Oncotarget* 7(30):47593–47608
 41. Yang L, Zhang X, Li H, Liu J (2016) The long noncoding RNA HOTAIR activates autophagy by upregulating ATG3 and ATG7 in hepatocellular carcinoma. *Mol BioSyst* 12:2605–2612
 42. Yu J, Han J, Zhang J, Li G, Liu H, Cui X, Xu Y, Li T, Liu J, Wang C (2016) The long noncoding RNAs PVT1 and uc002mbe.2 In sera provide a new supplementary method for hepatocellular carcinoma diagnosis. *Medicine (Baltimore)* 95:e4436
 43. Yuan P, Cao W, Zang Q, Li G, Guo X, Fan J (2016) The HIF-2 α -MALAT1-miR-216b axis regulates multi-drug resistance of hepatocellular carcinoma cells via modulating autophagy. *Biochem Biophys Res Commun* 478(3):1067–1073
 44. Zhou N, Si Z, Li T, Chen G, Zhang Z, Qi H (2016) Long non-coding RNA CCAT2 functions as an oncogene in hepatocellular

- carcinoma, regulating cellular proliferation, migration and apoptosis. *Oncol Lett* 12:132–138
45. Zhu XT, Yuan JH, Zhu TT, Li YY, Cheng XY (2016) Long noncoding RNA GPC3-AS1 promotes hepatocellular carcinoma progression via epigenetically activating GPC3. *FEBS J* 283(20):3739–3754
 46. Chauhan R, Lahiri N (2016) Tissue- and serum-associated biomarkers of hepatocellular carcinoma. *Biomark Cancer* 8:37–55
 47. Liu YR, Tang RX, Huang WT, Ren FH, He RQ, Yang LH, Luo DZ, Dang YW, Chen G (2015) Long noncoding RNAs in hepatocellular carcinoma: novel insights into their mechanism. *World J Hepatol* 7:2781–2791
 48. Shi L, Peng F, Tao Y, Fan X, Li N (2016) Roles of long noncoding RNAs in hepatocellular carcinoma. *Virus Res* 223:131–139
 49. Yang X, Xie X, Xiao YF, Xie R, Hu CJ, Tang B, Li BS, Yang SM (2015) The emergence of long non-coding RNAs in the tumorigenesis of hepatocellular carcinoma. *Cancer Lett* 360:119–124
 50. Panzitt K, Tschernatsch MM, Guelly C, Moustafa T, Stradner M, Strohmaier HM, Buck CR, Denk H, Schroeder R, Trauner M, Zatloukal K (2007) Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. *Gastroenterology* 132:330–342
 51. Li SP, Xu HX, Yu Y, He JD, Wang Z, Xu YJ, Wang CY, Zhang HM, Zhang RX, Zhang JJ, Yao Z, Shen ZY (2016) LncRNA HULC enhances epithelial-mesenchymal transition to promote tumorigenesis and metastasis of hepatocellular carcinoma via the miR-200a-3p/ZEB1 signaling pathway. *Oncotarget* 7 (27):42431–42446
 52. Xie H, Ma H, Zhou D (2013) Plasma HULC as a promising novel biomarker for the detection of hepatocellular carcinoma. *Biomed Res Int* 2013:136106
 53. Wang J, Liu X, Wu H, Ni P, Gu Z, Qiao Y, Chen N, Sun F, Fan Q (2010) CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic Acids Res* 38:5366–5383
 54. Di Bisceglie AM (2009) Hepatitis B and hepatocellular carcinoma. *Hepatology* 49: S56–S60
 55. Kim CM, Koike K, Saito I, Miyamura T, Jay G (1991) HBx gene of hepatitis B virus induces liver cancer in transgenic mice. *Nature* 351:317–320
 56. Muroyama R, Kato N, Yoshida H, Otsuka M, Moriyama M, Wang Y, Shao RX, Dharel N, Tanaka Y, Ohta M, Tateishi R, Shiina S, Tatsukawa M, Fukai K, Imazeki F, Yokosuka O, Shiratori Y, Omata M (2006) Nucleotide change of codon 38 in the X gene of hepatitis B virus genotype C is associated with an increased risk of hepatocellular carcinoma. *J Hepatol* 45:805–812
 57. Du Y, Kong G, You X, Zhang S, Zhang T, Gao Y, Ye L, Zhang X (2012) Elevation of highly up-regulated in liver cancer (HULC) by hepatitis B virus X protein promotes hepatoma cell proliferation via down-regulating p18. *J Biol Chem* 287:26302–26311
 58. Hammerle M, Gutschner T, Uckelmann H, Ozgur S, Fiskin E, Gross M, Skawran B, Geffers R, Longerich T, Brehahn K, Schirmacher P, Stoecklin G, Diederichs S (2013) Posttranscriptional destabilization of the liver-specific long noncoding RNA HULC by the IGF2 mRNA-binding protein 1 (IGF2BP1). *Hepatology* 58:1703–1712
 59. Lu Z, Xiao Z, Liu F, Cui M, Li W, Yang Z, Li J, Ye L, Zhang X (2016) Long non-coding RNA HULC promotes tumor angiogenesis in liver cancer by up-regulating sphingosine kinase 1 (SPHK1). *Oncotarget* 7:241–254
 60. Alvarez SE, Harikumar KB, Hait NC, Allegood J, Strub GM, Kim EY, Maceyka M, Jiang H, Luo C, Kordula T, Milstien S, Spiegel S (2010) Sphingosine-1-phosphate is a missing cofactor for the E3 ubiquitin ligase TRAF2. *Nature* 465:1084–1088
 61. Liu Y, Deng J, Wang L, Lee H, Armstrong B, Scuto A, Kowolik C, Weiss LM, Forman S, Yu H (2012) S1PR1 is an effective target to block STAT3 signaling in activated B cell-like diffuse large B-cell lymphoma. *Blood* 120:1458–1465
 62. Nagahashi M, Ramachandran S, Kim EY, Allegood JC, Rashid OM, Yamada A, Zhao R, Milstien S, Zhou H, Spiegel S, Takabe K (2012) Sphingosine-1-phosphate produced by sphingosine kinase 1 promotes breast cancer progression by stimulating angiogenesis and lymphangiogenesis. *Cancer Res* 72:726–735
 63. Tiraboschi P, Hansen LA, Thal LJ, Corey-Bloom J (2004) The importance of neuritic plaques and tangles to the development and evolution of AD. *Neurology* 62:1984–1989
 64. Murphy MP, LeVine H 3rd (2010) Alzheimer's Disease and the amyloid-beta peptide. *J Alzheimers Dis* 19:311–323
 65. Prince MWA, Guerchet M, Ali GC, Wu YT, Prina M, Alzheimer's Disease International

- (2015). World Alzheimer Report: The Global Impact of Dementia
66. Querfurth HW, LaFerla FM (2010) Alzheimer's Disease. *N Engl J Med* 362:329–344
 67. Todd S, Barr S, Roberts M, Passmore AP (2013) Survival in dementia and predictors of mortality: a review. *Int J Geriatr Psychiatry* 28:1109–1124
 68. Lin D, Pestova TV, Hellen CU, Tiedge H (2008) Translational control by a small RNA: dendritic BCL RNA targets the eukaryotic initiation factor 4A helicase mechanism. *Mol Cell Biol* 28:3008–3019
 69. Lukiw WJ, Handley P, Wong L, Crapper McLachlan DR (1992) BC200 RNA in normal human neocortex, non-Alzheimer dementia (NAD), and senile dementia of the Alzheimer type (AD). *Neurochem Res* 17:591–597
 70. Mus E, Hof PR, Tiedge H (2007) Dendritic BC200 RNA in aging and in Alzheimer's disease. *Proc Natl Acad Sci U S A* 104:10679–10684
 71. Bussiere T, Gold G, Kovari E, Giannakopoulos P, Bouras C, Perl DP, Morrison JH, Hof PR (2003) Stereologic analysis of neurofibrillary tangle formation in prefrontal cortex area 9 in aging and Alzheimer's disease. *Neuroscience* 117:577–592
 72. Engstrom PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, Lavorgna G, Brozzi A, Luzzi L, Tan SL, Yang L, Kunarso G, Ng EL, Batalov S, Wahlestedt C, Kai C, Kawai J, Carninci P, Hayashizaki Y, Wells C, Bajic VB, Orlando V, Reid JF, Lenhard B, Lipovich L (2006) Complex loci in human and mouse genomes. *PLoS Genet* 2:e47
 73. Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, Morgan TE, Finch CE, St Laurent G 3rd, Kenny PJ, Wahlestedt C (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat Med* 14:723–730
 74. Faghihi MA, Zhang M, Huang J, Modarresi F, Van der Brug MP, Nalls MA, Cookson MR, St-Laurent G 3rd, Wahlestedt C (2010) Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome Biol* 11:R56
 75. Kang MJ, Abdelmohsen K, Hutchison ER, Mitchell SJ, Grammatikakis I, Guo R, Noh JH, Martindale JL, Yang X, Lee EK, Faghihi MA, Wahlestedt C, Troncoso JC, Pletnikova O, Perrone-Bizzozero N, Resnick SM, de Cabo R, Mattson MP, Gorospe M (2014) HuD regulates coding and noncoding RNA to induce APP→Aβeta processing. *Cell Rep* 7:1401–1409
 76. Ciarlo E, Massone S, Penna I, Nizzari M, Gigoni A, Dieci G, Russo C, Florio T, Cancedda R, Pagano A (2013) An intronic ncRNA-dependent regulation of SORL1 expression affecting Aβeta formation is upregulated in post-mortem Alzheimer's disease brain samples. *Dis Model Mech* 6:424–433
 77. Massone S, Ciarlo E, Vella S, Nizzari M, Florio T, Russo C, Cancedda R, Pagano A (2012) NDM29, A RNA polymerase III-dependent non coding RNA, promotes amyloidogenic processing of APP and amyloid beta secretion. *Biochim Biophys Acta* 1823:1170–1177
 78. Massone S, Vassallo I, Fiorino G, Castelnuovo M, Barbieri F, Borghi R, Tabaton M, Robello M, Gatta E, Russo C, Florio T, Dieci G, Cancedda R, Pagano A (2011) 17A, A novel non-coding RNA, regulates GABA B alternative splicing and signaling in response to inflammatory stimuli and in Alzheimer disease. *Neurobiol Dis* 41:308–317
 79. Lee DY, Moon J, Lee ST, Jung KH, Park DK, Yoo JS, Sunwoo JS, Byun JI, Shin JW, Jeon D, Jung KY, Kim M, Lee SK, Chu K (2015) Distinct expression of long non-coding RNAs in an Alzheimer's disease model. *J Alzheimers Dis* 45:837–849
 80. Yang B, Xia ZA, Zhong B, Xiong X, Sheng C, Wang Y, Gong W, Cao Y, Wang Z, Peng W (2016) Distinct hippocampal expression profiles of long non-coding RNAs in an Alzheimer's disease model. *Mol Neurobiol* 54(7):4833–4846
 81. Zhou X, Xu J (2015) Identification of Alzheimer's disease-associated long noncoding RNAs. *Neurobiol Aging* 36:2925–2931
 82. Magistri M, Velmeshev D, Makhmutova M, Faghihi MA (2015) Transcriptomics profiling of Alzheimer's disease reveal neurovascular defects, altered amyloid-beta homeostasis, and deregulated expression of long noncoding RNAs. *J Alzheimers Dis* 48:647–665
 83. Sun X, Wong D (2016) Long non-coding RNA-mediated regulation of glucose homeostasis and diabetes. *Am J Cardiovasc Dis* 6:17–25
 84. Rachmilewitz J, Goshen R, Ariel I, Schneider T, de Groot N, Hochberg A (1992) Parental imprinting of the human H19 gene. *FEBS Lett* 309:25–28
 85. Gabory A, Jammes H, Dandolo L (2010) The H19 locus: role of an imprinted non-coding

- RNA in growth and development. *BioEssays* 32:473–480
86. Raveh E, Matouk IJ, Gilon M, Hochberg A (2015) The H19 long non-coding RNA in cancer initiation, progression and metastasis—a proposed unifying theory. *Mol Cancer* 14:184
 87. Leighton PA, Ingram RS, Eggenschwiler J, Efstratiadis A, Tilghman SM (1995) Disruption of imprinting caused by deletion of the H19 gene region in mice. *Nature* 375:34–39
 88. Petry CJ, Evans ML, Wingate DL, Ong KK, Reik W, Constanca M, Dunger DB (2010) Raised late pregnancy glucose concentrations in mice carrying pups with targeted disruption of H19delta13. *Diabetes* 59:282–286
 89. Petry CJ, Seear RV, Wingate DL, Acerini CL, Ong KK, Hughes IA, Dunger DB (2011) Maternally transmitted foetal H19 variants and associations with birth weight. *Hum Genet* 130:663–670
 90. Gao Y, Wu F, Zhou J, Yan L, Jurczak MJ, Lee HY, Yang L, Mueller M, Zhou XB, Dandolo L, Szendroedi J, Roden M, Flannery C, Taylor H, Carmichael GG, Shulman GI, Huang Y (2014) The H19/let-7 double-negative feedback loop contributes to glucose metabolism in muscle cells. *Nucleic Acids Res* 42:13799–13811
 91. Kallen AN, Zhou XB, Xu J, Qiao C, Ma J, Yan L, Lu L, Liu C, Yi JS, Zhang H, Min W, Bennett AM, Gregory RI, Ding Y, Huang Y (2013) The imprinted H19 lncRNA antagonizes let-7 microRNAs. *Mol Cell* 52:101–112
 92. Miyoshi N, Wagatsuma H, Wakana S, Shiroishi T, Nomura M, Aisaka K, Kohda T, Surani MA, Kaneko-Ishino T, Ishino F (2000) Identification of an imprinted gene, *Meg3/Gtl2* and its human homologue *MEG3*, first mapped on mouse distal chromosome 12 and human chromosome 14q. *Genes Cells* 5:211–220
 93. Guo Q, Qian Z, Yan D, Li L, Huang L (2016) LncRNA-MEG3 inhibits cell proliferation of endometrial carcinoma by repressing notch signaling. *Biomed Pharmacother* 82:589–594
 94. Lu KH, Li W, Liu XH, Sun M, Zhang ML, Wu WQ, Xie WP, Hou YY (2013) Long non-coding RNA MEG3 inhibits NSCLC cells proliferation and induces apoptosis by affecting p53 expression. *BMC Cancer* 13:461
 95. Luo G, Wang M, Wu X, Tao D, Xiao X, Wang L, Min F, Zeng F, Jiang G (2015) Long non-coding RNA MEG3 inhibits cell proliferation and induces apoptosis in prostate cancer. *Cell Physiol Biochem* 37:2209–2220
 96. Zhu X, Wu YB, Zhou J, Kang DM (2016) Upregulation of lncRNA MEG3 promotes hepatic insulin resistance via increasing FoxO1 expression. *Biochem Biophys Res Commun* 469:319–325
 97. You L, Wang N, Yin D, Wang L, Jin F, Zhu Y, Yuan Q, De W (2016) Downregulation of long noncoding RNA *Meg3* affects insulin synthesis and secretion in mouse pancreatic beta cells. *J Cell Physiol* 231:852–862
 98. Moran I, Akerman I, van de Bunt M, Xie R, Benazra M, Nammo T, Arnes L, Nakic N, Garcia-Hurtado J, Rodriguez-Segui S, Pasquali L, Sauty-Colace C, Beucher A, Scharfmann R, van Arensbergen J, Johnson PR, Berry A, Lee C, Harkins T, Gmyr V, Pattou F, Kerr-Conte J, Piemonti L, Berney T, Hanley N, Gloyn AL, Sussel L, Langman L, Brayman KL, Sander M, McCarthy MI, Ravassard P, Ferrer J (2012) Human beta cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metab* 16:435–448
 99. Cho YS, Chen CH, Hu C, Long J, Ong RT, Sim X, Takeuchi F, Wu Y, Go MJ, Yamauchi T, Chang YC, Kwak SH, Ma RC, Yamamoto K, Adair LS, Aung T, Cai Q, Chang LC, Chen YT, Gao Y, Hu FB, Kim HL, Kim S, Kim YJ, Lee JJ, Lee NR, Li Y, Liu JJ, Lu W, Nakamura J, Nakashima E, Ng DP, Tay WT, Tsai FJ, Wong TY, Yokota M, Zheng W, Zhang R, Wang C, So WY, Ohnaka K, Ikegami H, Hara K, Cho YM, Cho NH, Chang TJ, Bao Y, Hedman AK, Morris AP, McCarthy MI, Takayanagi R, Park KS, Jia W, Chuang LM, Chan JC, Maeda S, Kadowaki T, Lee JY, Wu JY, Teo YY, Tai ES, Shu XO, Mohlke KL, Kato N, Han BG, Seielstad M (2012) Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat Genet* 44:67–72
 100. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, Wheeler E, Glazer NL, Bouatia-Naji N, Gloyn AL, Lindgren CM, Magi R, Morris AP, Randall J, Johnson T, Elliott P, Rybin D, Thorleifsson G, Steinthorsdottir V, Henneman P, Grallert H, Dehghan A, Hot-tenga JJ, Franklin CS, Navarro P, Song K, Goel A, Perry JR, Egan JM, Lajunen T, Grarup N, Sparso T, Doney A, Voight BF, Stringham HM, Li M, Kanoni S, Shrader P, Cavalcanti-Proenca C, Kumari M, Qi L, Timpson NJ, Gieger C, Zabena C, Rocheleau G, Ingelsson E, An P, O'Connell J, Luan J, Elliott A, McCarroll

- SA, Payne F, Roccacaccia RM, Pattou F, Sethupathy P, Ardlie K, Ariyurek Y, Balkau B, Barter P, Beilby JP, Ben-Shlomo Y, Benediktsson R, Bennett AJ, Bergmann S, Bochud M, Boerwinkle E, Bonnefond A, Bonnycastle LL, Borch-Johnsen K, Bottcher Y, Brunner E, Bumpstead SJ, Charpentier G, Chen YD, Chines P, Clarke R, Coin LJ, Cooper MN, Cornelis M, Crawford G, Crisponi L, Day IN, de Geus EJ, Delplanque J, Dina C, Erdos MR, Fedson AC, Fischer-Rosinsky A, Forouhi NG, Fox CS, Frants R, Franzosi MG, Galan P, Goodarzi MO, Graessler J, Groves CJ, Grundy S, Gwilliam R, Gyllenstein U, Hadjadj S, Hallmans G, Hammond N, Han X, Hartikainen AL, Hassanali N, Hayward C, Heath SC, Hercberg S, Herder C, Hicks AA, Hillman DR, Hingorani AD, Hofman A, Hui J, Hung J, Isomaa B, Johnson PR, Jorgensen T, Julia A, Kaakinen M, Kaprio J, Kesaniemi YA, Kivimaki M, Knight B, Koskinen S, Kovacs P, Kyvik KO, Lathrop GM, Lawlor DA, Le Bacquer O, Lecoecur C, Li Y, Lyssenko V, Mahley R, Mangino M, Manning AK, Martinez-Larrad MT, McAteer JB, McCulloch LJ, McPherson R, Meisinger C, Melzer D, Meyre D, Mitchell BD, Morken MA, Mukherjee S, Naitza S, Narisu N, Neville MJ, Oostra BA, Orru M, Pakyz R, Palmer CN, Paolisso G, Pattaro C, Pearson D, Peden JF, Pedersen NL, Perola M, Pfeiffer AF, Pichler I, Polasek O, Posthuma D, Potter SC, Pouta A, Province MA, Psaty BM, Rathmann W, Rayner NW, Rice K, Ripatti S, Rivadeneira F, Roden M, Rolandsson O, Sandbaek A, Sandhu M, Sanna S, Sayer AA, Scheet P, Scott LJ, Seedorf U, Sharp SJ, Shields B, Sigurdsson G, Sijbrands EJ, Silveira A, Simpson L, Singleton A, Smith NL, Sovio U, Swift A, Syddall H, Syvanen AC, Tanaka T, Thorand B, Tichet J, Tonjes A, Tuomi T, Uitterlinden AG, van Dijk KW, van Hoek M, Varma D, Visvikis-Siest S, Vitart V, Vogelzangs N, Waeber G, Wagner PJ, Walley A, Walters GB, Ward KL, Watkins H, Weedon MN, Wild SH, Willemsen G, Wittteman JC, Yarnell JW, Zeggini E, Zelenika D, Zethelius B, Zhai G, Zhao JH, Zillikens MC, Borecki IB, Loos RJ, Meneton P, Magnusson PK, Nathan DM, Williams GH, Hattersley AT, Silander K, Salomaa V, Smith GD, Bornstein SR, Schwarz P, Spranger J, Karpe F, Shuldiner AR, Cooper C, Dedoussis GV, Serrano-Rios M, Morris AD, Lind L, Palmer LJ, Hu FB, Franks PW, Ebrahim S, Marmot M, Kao WH, Pankow JS, Sampson MJ, Kuusisto J, Laakso M, Hansen T, Pedersen O, Pramstaller PP, Wichmann HE, Illig T, Rudan I, Wright AF, Stumvoll M, Campbell H, Wilson JF, Bergman RN, Buchanan TA, Collins FS, Mohlke KL, Tuomilehto J, Valle TT, Altshuler D, Rotter JI, Siscovick DS, Penninx BW, Boomsma DI, Deloukas P, Spector TD, Frayling TM, Ferrucci L, Kong A, Thorsteinsdottir U, Stefansson K, van Duijn CM, Aulchenko YS, Cao A, Scuteri A, Schlessinger D, Uda M, Ruukonen A, Jarvelin MR, Waterworth DM, Vollenweider P, Peltonen L, Mooser V, Abecasis GR, Wareham NJ, Sladek R, Froguel P, Watanabe RM, Meigs JB, Groop L, Boehnke M, McCarthy MI, Florez JC, Barroso I (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* 42:105–116
101. Kooner JS, Saleheen D, Sim X, Sehmi J, Zhang W, Frossard P, Been LF, Chia KS, Dimas AS, Hassanali N, Jafar T, Jowett JB, Li X, Radha V, Rees SD, Takeuchi F, Young R, Aung T, Basit A, Chidambaram M, Das D, Grundberg E, Hedman AK, Hydrie ZI, Islam M, Khor CC, Kowlessur S, Kristensen MM, Liju S, Lim WY, Matthews DR, Liu J, Morris AP, Nica AC, Pinidiyapathirage JM, Prokopenko I, Rasheed A, Samuel M, Shah N, Shera AS, Small KS, Suo C, Wickremasinghe AR, Wong TY, Yang M, Zhang F, Abecasis GR, Barnett AH, Caulfield M, Deloukas P, Frayling TM, Froguel P, Kato N, Katulanda P, Kelly MA, Liang J, Mohan V, Sanghera DK, Scott J, Seielstad M, Zimmet PZ, Elliott P, Teo YY, McCarthy MI, Danesh J, Tai ES, Chambers JC (2011) Genome-wide association study in individuals of south Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet* 43:984–989
102. Strawbridge RJ, Dupuis J, Prokopenko I, Barker A, Ahlqvist E, Rybin D, Petrie JR, Travers ME, Bouatia-Naji N, Dimas AS, Nica A, Wheeler E, Chen H, Voight BF, Taneera J, Kanoni S, Peden JF, Turrini F, Gustafsson S, Zabena C, Almgren P, Barker DJ, Barnes D, Dennison EM, Eriksson JG, Eriksson P, Eury E, Folkersen L, Fox CS, Frayling TM, Goel A, Gu HF, Horikoshi M, Isomaa B, Jackson AU, Jameson KA, Kajantie E, Kerr-Conte J, Kuulasmaa T, Kuusisto J, Loos RJ, Luan J, Makrillakis K, Manning AK, Martinez-Larrad MT, Narisu N, Nastase Mannila M, Ohrvik J, Osmond C, Pascoe L, Payne F, Sayer AA, Sennblad B, Silveira A, Stancakova A, Stirrups K, Swift AJ, Syvanen AC, Tuomi T, van 't Hooft FM, Walker M, Weedon MN,

- Xie W, Zethelius B, Ongen H, Malarstig A, Hopewell JC, Saleheen D, Chambers J, Parish S, Danesh J, Kooner J, Ostenson CG, Lind L, Cooper CC, Serrano-Rios M, Ferrannini E, Forsen TJ, Clarke R, Franzosi MG, Seedorf U, Watkins H, Froguel P, Johnson P, Deloukas P, Collins FS, Laakso M, Dermizakis ET, Boehnke M, McCarthy MI, Wareham NJ, Groop L, Pattou F, Gloyn AL, Dedoussis GV, Lyssenko V, Meigs JB, Barroso I, Watanabe RM, Ingelsson E, Langenberg C, Hamsten A, Florez JC (2011) Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* 60:2624–2634
103. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segre AV, van Hoek M, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, Boerwinkle E, Bonnycastle LL, Bengtsson Bostrom K, Bravenboer B, Bumpstead S, Burt NP, Charpentier G, Chines PS, Cornelis M, Couper DJ, Crawford G, Doney AS, Elliott KS, Elliott AL, Erdos MR, Fox CS, Franklin CS, Ganser M, Gieger C, Grarup N, Green T, Griffin S, Groves CJ, Guiducci C, Hadjadj S, Hassanali N, Herder C, Isomaa B, Jackson AU, Johnson PR, Jorgensen T, Kao WH, Klopp N, Kong A, Kraft P, Kuusisto J, Lauritzen T, Li M, Lieveise A, Lindgren CM, Lyssenko V, Marre M, Meitinger T, Midthjell K, Morken MA, Narisu N, Nilsson P, Owen KR, Payne F, Perry JR, Petersen AK, Platou C, Proenca C, Prokopenko I, Rathmann W, Rayner NW, Robertson NR, Rocheleau G, Roden M, Sampson MJ, Saxena R, Shields BM, Shradler P, Sigurdsson G, Sparso T, Strassburger K, Stringham HM, Sun Q, Swift AJ, Thorand B, Tichet J, Tuomi T, van Dam RM, van Haeften TW, van Herpt T, van Vliet-Ostaptchouk JV, Walters GB, Weedon MN, Wijmenga C, Witteman J, Bergman RN, Cauchi S, Collins FS, Gloyn AL, Gyllenstein U, Hansen T, Hide WA, Hitman GA, Hofman A, Hunter DJ, Hveem K, Laakso M, Mohlke KL, Morris AD, Palmer CN, Pramstaller PP, Rudan I, Sijbrands E, Stein LD, Tuomilehto J, Uitterlinden A, Walker M, Wareham NJ, Watanabe RM, Abecasis GR, Boehm BO, Campbell H, Daly MJ, Hattersley AT, Hu FB, Meigs JB, Pankow JS, Pedersen O, Wichmann HE, Barroso I, Florez JC, Frayling TM, Groop L, Sladek R, Thorsteinsdottir U, Wilson JF, Illig T, Froguel P, van Duijn CM, Stefansson K, Altshuler D, Boehnke M, McCarthy MI (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 42:579–589
104. Fadista J, Vikman P, Laakso EO, Mollet IG, Esguerra JL, Taneera J, Storm P, Osmark P, Ladenvall C, Prasad RB, Hansson KB, Finotello F, Uvebrant K, Ofori JK, Di Camillo B, Krus U, Cilio CM, Hansson O, Eliasson L, Rosengren AH, Renstrom E, Wollheim CB, Groop L (2014) Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc Natl Acad Sci U S A* 111:13924–13929
105. Pasmant E, Sabbagh A, Vidaud M, Bieche I (2011) ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J* 25:444–448
106. Kotake Y, Nakagawa T, Kitagawa K, Suzuki S, Liu N, Kitagawa M, Xiong Y (2011) Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15 (INK4B) tumor suppressor gene. *Oncogene* 30:1956–1962
107. Pullen TJ, Rutter GA (2013) Could lncRNAs contribute to beta-cell identity and its loss in type 2 diabetes? *Biochem Soc Trans* 41:797–801

Part II

Methods for Gene Identification

Identification of Disease-Related Genes Using a Genome-Wide Association Study Approach

Tobias Wohland and Dorit Schleinitz

Abstract

Genome-wide association studies (GWAS) provide a hypothesis-free approach to discover genetic variants contributing to the risk of a certain disease or disease-related trait. Ongoing efforts to annotate the human genome have helped to localize disease-causing variants and point to mechanisms by which genetic variants might exert functional effects. By integrating bioinformatics approaches with in vivo and in vitro genomic strategies to predict and subsequently validate the functional roles of GWAS-identified variants, disease-related pathways can be characterized, providing new possibilities for therapeutic intervention. Here, we describe a basic workflow, from sample preparation to data analysis, for performing a GWAS to identify disease genes. We also discuss resources for the annotation and interpretation of GWAS results.

Key words GWAS, Affymetrix, Illumina, R, GenABEL, SNP annotation

1 Introduction

1.1 General Overview of GWAS

A central aim in human genetics research is to identify DNA variants that contribute to disease [1]. Genome-wide association studies (GWAS) analyze DNA sequence variations, primarily single nucleotide polymorphisms (SNPs), spanning the genome to identify genetic risk factors for specific diseases [2]. In contrast to linkage studies, which rely on segregation of alleles within families, no prior information on relatedness is required in GWAS [3]. Instead, association analyses compare allele frequencies of markers with phenotypes in a population using a case-control study design, which not only eases the burden of sample collection, but also utilizes simpler analytical methods than those used in family-based linkage studies. Without doubt, technological progress, large genome sequencing projects (*Human Genome Project*, *1000 Genomes*), and the development of advanced bioinformatics tools have allowed GWAS to evolve into a powerful approach for investigating the genetic architecture of human disease. At the time of this writing, the *GWAS Catalog* (www.ebi.ac.uk/gwas/), which

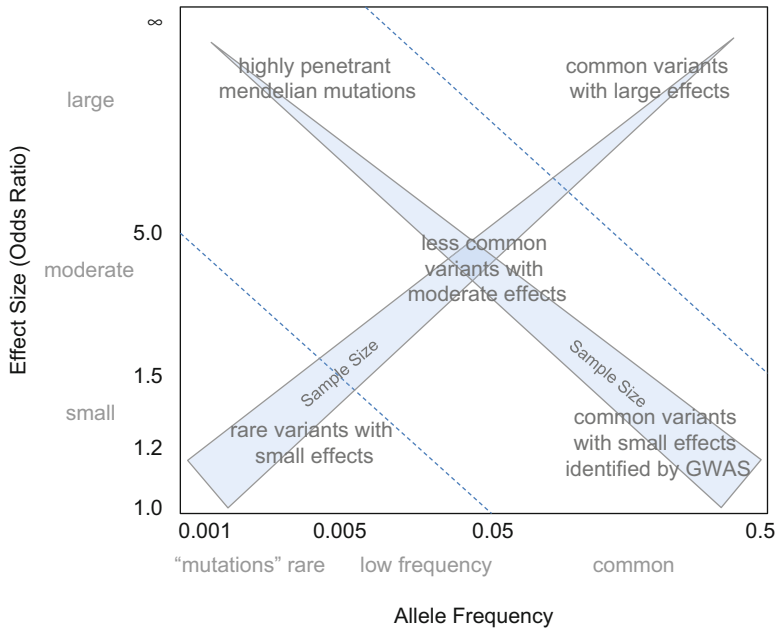


Fig. 1 Spectrum of allele effects associated with disease. Mendelian disorders are characterized by rare, highly penetrant alleles with large effect sizes, which can be identified with a small sample size. GWAS findings are often represented by associations of common variants with small effect sizes identified in large sample sets. *Blue arrows* represent the required sample size, where the top of the arrow represents a small sample size and the base a large sample size. Adopted and edited from Bush WS & Moore JH (2012) PLoS Computational Biology 8(12):e1002822 (see **Note 1**)

is the largest resource for published GWAS, listing all studies meeting the criteria of (1) assaying >100,000 SNPs, and (2) $p < 1.0 \times 10^{-5}$, contained 2634 studies and 29,592 unique SNP–trait associations [4, 5].

Allele frequency and effect size are the two dimensions used to conceptualize disease associations [2]. Mendelian disorders are typically characterized by rare, highly penetrant alleles with large effect sizes. In contrast, GWAS findings are often represented by associations of common variants with small effect sizes (Fig. 1). The smaller the effect of the allele on the disease or trait of interest, the larger the sample size required to detect the differences. The linkage disequilibrium (LD) pattern in the human genome both simplifies and complicates the analyses and interpretation of GWAS results. On the one hand, SNPs can be genotyped as surrogates for other variants contained in one LD group, thereby reducing the number of markers to be assayed. On the other hand, however, the lead SNP showing the strongest evidence for association is rarely the actual causal variant, which is usually tagged by indirect association [2]. In addition, long-range effects of regulatory active SNPs will not always tag (only) the nearest gene. For example, variants within introns of the fat mass and obesity-associated gene (*FTO*)

have long been reproducibly associated with increased risk for obesity and type 2 diabetes in GWAS. However, while studies in mice demonstrated that *FTO* expression levels influence body mass and composition phenotypes, no direct connection between the obesity-associated variants and *FTO* expression or function have ever been made [6]. Instead, the obesity-associated *FTO* region was found to directly interact with the promoters of *FTO* and a nearby gene, Iroquois-class homeodomain protein (*IRX3*), in human, mouse, and zebrafish genomes [6]. Long-range enhancers within this region also recapitulated aspects of *IRX3* expression. In human brains, *FTO* SNPs were associated with expression of *IRX3*, but not *FTO*. Body-weight reduction by loss of fat mass was observed in *Irx3*-deficient mice directly linking *IRX3* expression with regulation of body mass and composition [6].

Once a region of association has been identified, bioinformatics approaches and functional assays become essential for pinpointing the causal variant or the affected gene, and determining the mechanism by which the variant(s) exerts its effects. Genetic factors to predict personal disease risk and identify the biological underpinnings of disease susceptibility may help to develop new prevention and treatment strategies [2]. Nevertheless, the initial generation of genome-wide SNP information per sample remains costly. However, given the fact that study participants are typically well phenotyped, such approaches have continuing benefit for serving in secondary GWAS (i.e., testing for association with traits other than the primary study) and participating in large, multicenter consortia.

For high-density genotyping, two types of arrays are primarily used: the Illumina[®] BeadArray for SNP genotyping and the Affymetrix[®] Axiom[®] Genome-Wide Array. The protocol described here will refer to these arrays in a general survey, while the main focus will be on the data handling and bioinformatics analysis (i.e., the GenABEL workflow). Because every data-set and subsequent analysis have their own difficulties, this chapter may not cover issues specific to certain circumstances. In light of this, we have prepared this chapter to allow an investigator to brachiate through the single steps to perform an individualized analysis. This chapter therefore seeks to provide a basic, overall workflow, from sample preparation to data analysis, for performing a GWAS to identify disease genes. We also discuss resources for SNP annotation and interpretation of GWAS results for subsequent downstream analyses.

1.2 Considerations in Preparing for GWAS

1.2.1 Study Design

To achieve a meaningful result in a GWAS, a thoughtful approach to characterize the phenotype of interest is critical (*see Note 1*) [2]. Phenotypes are classified as categorical (often binary, case-control) or quantitative (e.g., anthropometric or metabolic measures). From a statistical point of view, quantitative, disease-related traits (i.e., endophenotypes) are preferred because they improve

power to detect a genetic effect and often have a more interpretable outcome. For example, a quantitative trait can represent a units-change of the trait per allele or genotype class [2]. However, not all disease traits have well-established measures; in this case, individuals are usually classified as either affected or unaffected. Thus, while quantitative outcomes are preferred, they are not required for a successful study [2, 7].

1.2.2 Standardized Phenotype Criteria

The definition of rigorous phenotype criteria is essential, and particularly critical for multicenter studies to prevent the introduction of site-based effects into the study [2]. However, variability in phenotypes based on factual use of those criteria by the clinicians, or on biased measurement, may not be completely eliminated. Therefore, statistical analyses are going to be adjusted whereby the participating study centers are coded into a categorical variable and used as a covariate.

1.2.3 Power Calculations

Power calculations are not only important for optimizing study design, but are also critical for ensuring meaningful results (*see Note 2*) [8]. Power calculations go hand-in-hand with sample size. Many aspects of study design, such as selection of subjects, definition and measurement of phenotype, choice of how many and which genetic variants to analyze, inclusion of covariates and other possible confounding factors, and the statistical method to be used, can be controlled by the researchers [8]. It is always worthwhile to maximize the statistical power of a study, given the constraints imposed by nature and limitations in resources (*see Note 2*) [8, 9]. The power or sensitivity of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis (H_0) when the alternative hypothesis (H_1) is true. Tools for power and sample size calculations, such as the G*power program [10, 11] or Quanto (<http://biostats.usc.edu/Quanto.html>), are freely available online (*see Note 3*).

2 Materials and User Guides

The genome-wide genotyping wet lab protocols require specific equipment, plastics, and reagents. To list them here would go beyond the scope of this paper, which instead focuses on data handling and analysis. This section thus provides information for resources where protocols and equipment for performing a GWAS can be found.

2.1 Target DNA

1. Depending on the platform and chip type used, 200–500 ng purified DNA per sample is recommended for genotyping.

2.2 Genotyping Protocols

1. The “Axiom[®] 2.0 Assay Manual Workflow User Guide” is available at <http://www.affymetrix.com/support/technical/manuals.affx>. The guide is set up for 96 samples; 24 sample- and 384 sample- versions are also available (*see Note 4*).
2. The “Infinium LCG Quad Assay Guide” for e.g., the Illumina[®] Infinium Omni5Exome-4 BeadChip array (*see Note 4*) is available at http://support.illumina.com/array/array_kits/infinium_humanomni5exome_beadchip_kit.html (*see Note 5*).

2.3 Lab Equipment, Plastics, and Reagents

1. Equipment and supplies for the Axiom[®] 2.0 Assay for 96 Samples are listed in the Site Preparation Guide; additional equipment and consumables are listed in the protocol (*see Subheading 2.2, item 1*).
2. The “Infinium assay lab setup and procedures” manual is available at http://support.illumina.com/array/array_kits/infinium_humanomni5exome_beadchip_kit/documentation.html and includes all general recommendations for lab equipment, plastics, and reagents. Additional equipment specific for the Infinium Omni5Exome-4 BeadChip array is listed in the manual given in Subheading 2.2, **item 2**.

2.4 Instruments, Operating System, and Software

2.4.1 Instruments

Table 1 lists the main instruments and software needed to perform genome-wide SNP genotyping and analyze the data. For further information and additional equipment, please refer to the manufacturer manuals (Subheading 2.3). More details regarding the operating system and the data analysis software are shown below.

2.4.2 Operating System (OS)

Because many users use Microsoft Windows, we provide a description based on the Microsoft OS, version 7 or newer. In general, we recommend using a computer with Linux as OS, because exploring data is much easier with the Linux bash shell and preserves computer resources like working memory. However, all commands displayed will work on both OS, as well as the Mac OS. Cygwin (<http://cygwin.com>) is a Linux Bash Shell clone with similar functionality. While working with a big text file, which is the case for almost all genetic formats, it is absolutely necessary to have the ability to explore the files in a fast way. The Bash Shell in Linux is the perfect tool for doing this, and Cygwin is a good copy for Windows (Table 2).

1. *Cygwin* has its own environment on the computer. Therefore, the data need to be copied to Cygwin prior to analysis. In the Cygwin-folder, there will be a home-folder. We recommend creating a username- and then a data-folder within that directory. Eventually, one should create a structure as follows:

```
C:\cygwin64\home\user\Data>
```

The next few lines display some commands, which are helpful with big text files:

Command ^a	Description
head <i>filename</i> ^b	Displays the first 10 lines of the file (one line of the file is potentially on multiple lines of the screen in case the file has more columns than can be displayed by the monitor)
head -nx <i>filename</i>	Displays the first x lines of the file (same limitation as for ‘head filename’)
head <i>filename</i> less -S	Displays the first 10 lines of the file (one line of the file is on one line of the screen—not all columns are displayed; depends on monitor width)
head <i>filename</i> cut -d “-fx	Displays column x of the first 10 lines of a space delimited file

^aFar more commands available; combining the different commands with a pipe (|) makes the Bash Shell an extremely powerful tool for bioinformaticians

^bReplace *filename* by the name of your file

2.4.3 Software

One major software package we will use extensively in this tutorial is R, an environment for statistical computing and graphics, which can be freely downloaded (Table 2) [12]. Because it is not possible to explain every single step, we strongly encourage the reader to become familiar with the syntax and logical behavior of R, which will make using the package easier. In general, selecting the appropriate software for performing a GWAS is crucial. While there are a number of software packages available, we recommend four widely used tools, including PLINK [13], GWASTools [14], SNPTEST2 [15], and the GenABEL-suite [16] (Table 2). In this protocol we will mostly use GenABEL, but provide a short overview of the other three programs below.

1. *PLINK* is a command-line based, open-source program [13], and is probably the most commonly used software for analysis of GWAS data. PLINK has a wide range of functions and is computationally efficient, because it is written in C/C++, one of the most powerful programming languages. A disadvantage of the native PLINK program is its limited flexibility. Due to the fact that one will stay in a kind of isolated environment, the maximum functionality will be reached at some point. However, extended PLINK has limited R-support, which is a strong advantage, as R is one of the major statistical environments. Regardless of whether PLINK is used or not, the data formats introduced by PLINK are very commonly used in the community and also by other software. For this reason, we will execute PLINK in the beginning of this tutorial (Subheading 3.2.2).

Table 1
Main instruments and software for the GWAS procedure


Genotyping		
	Affymetrix platform	Illumina Platform ^a
Instruments	GeneTitan [®] Multi-Channel Instrument	<ul style="list-style-type: none"> • Hybridization Oven • Water Circulator with Programmable Temperature Controls • HiScan System or iScan System
Software	GeneChip [®] Command Console [®] (AGCC) Axiom [™] Analysis Suite	Illumina [®] GenomeStudio Genotyping Module
Array	e.g., Axiom [®] Genome-Wide Population-Optimized Human Arrays	e.g., Illumina [®] Infinium Omni5Exome-4 BeadChip array
Data analysis		
	Component	Comment
Computer	Intel Core i3/AMD FX ^b 8GB RAM	Minimal requirements based on dataset of 500.000 markers and 1000 samples
Server	Intel Xeon/AMD Opteron ^b 8–12 CPU Cores 64GB RAM	Estimated for imputation and working with imputed data from the <i>1000 Genomes Project</i>
Operating system	Linux/Windows/Mac OS	Details on the operating system: Subheading 2.4.2
Software	PLINK Cygwin (only Windows) R/RStudio GenABEL	More information on the software please <i>see</i> Subheading 2.4.2

^aThe Illumina platform for SNP genotyping requires a set of specific items related to the handling of the samples, BeadChip arrays and reagents

^bRequirements are based on current CPU architecture but older CPU generations are also possible

2. *SNPTEST2* is another command line-based program, developed by the authors of *IMPUTE2*, a widely used software solution for imputing [15]. *SNPTEST2* was directly created for a workflow using imputed data based on imputation with *IMPUTE2*, which makes it easy to use with this kind of data. Since imputed data have a specific structure and require further editing before importing to other software, it is worthwhile to determine whether this program has value for the analytical needs of the reader.
3. *GWASTools* is a Bioconductor package that uses the R-language [14]. Because R is a statistical environment with a vast number of options for all kinds of data, use of *GWASTools* will increase the flexibility to edit and analyze your data. This is also true for *GenABEL*, which is also an R-package (*see Note 6*). A strong

Table 2
Links for user manuals and programs

Introduction to R	<ul style="list-style-type: none"> • http://cran.us.r-project.org/doc/manuals/R-intro.pdf • http://blog.revolutionanalytics.com/2013/08/google-video-r-tutorials.html
Install R	• https://cran.r-project.org/
Introduction to RStudio	• https://www.youtube.com/watch?v=5YmcEYTSN7k
Install RStudio	<ul style="list-style-type: none"> • https://www.rstudio.com/products/rstudio/download3/ → integrated development environment for R
Install Cygwin	• https://www.cygwin.com/ (only necessary for Windows)
Introduction to PLINK	• http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml
Install PLINK	<ul style="list-style-type: none"> • http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml <p>Usage note: → every PLINK command has to be typed in the windows command line tool (cmd) → to open the cmd: Press  → type “cmd” → press enter → to change the cmd to full-size mode type: mode 800 → press enter (command only necessary for Windows 7 and 8)</p>
Introduction to SNPTEST and SNPTEST2	• http://innovation.ox.ac.uk/licence-details/snptest/
Introduction to GenABEL	• http://www.genabel.org/tutorials
Install GenABEL	• https://cran.r-project.org/web/packages/GenABEL/index.html
Introduction to GWASTools	• http://bioconductor.org/packages/release/bioc/html/GWASTools.html

advantage of GWASTools is that it eliminates one of the major disadvantages of R, that is, the memory usage. Because everything in R, by default, is stored in the memory, problems can arise when dealing with big data on a local machine. This should be kept in mind when it comes to GWASTools.

4. *GenABEL*, or more specifically the GenABEL suite [16], is also an R-package, but unlike GWASTools, is not part of the Bioconductor project. The GenABEL suite has a comprehensive range of options for reading, editing, and analyzing data. Because of these options and its ease of use, we will focus on the use of GenABEL in this chapter. The bioinformatics analysis protocol given in this chapter (Subheading 3.3) was adopted from the official GenABEL-tutorial (Table 2) [17].

3 Methods

3.1 Wet Lab Procedures

3.1.1 DNA Preparation

For DNA extraction kits from Qiagen (e.g., QIAamp DNA Blood Kits) or Millipore are recommended as they have been tested by Affymetrix (please refer to the manufacturers handbooks). However, other kits may be used, as long as they avoid boiling or strong denaturants. DNA concentration and quality can be determined using NanoDrop™ spectrophotometry (Thermo Fisher Scientific) or PicoGreen (Thermo Fisher Scientific). In the event that the DNA preparation contain inhibitors (e.g., ethanol residuals, salt, proteins), which may disturb DNA amplification procedures, an additional cleanup step may be required (*see Note 7*).

3.1.2 Schematic Workflow for Affymetrix® and Illumina® SNP Arrays

The schematic workflow for the SNP arrays is shown in Fig. 2.

3.2 Bioinformatics Workflow

The analysis of genomic data, especially when increasing the informational power of a data-set with imputation, is a complex process where several steps must be considered (Fig. 3). While one can use specifically defined protocols for the wet lab part, available bioinformatics methods are much more diverse. For example, specific questions and hypotheses can require different ways to analyze the data. While it is not possible to include all options in a single chapter, we do provide a generalized pipeline below that should serve as a basic foundation upon which to build a study-specific bioinformatics flow.

3.2.1 Input Data

The input data for the GWAS depends on the output data from the genotyping-platform. As mentioned earlier, Illumina and Affymetrix are presently the two big players on the market. Both use self-developed software to analyze the raw output of the genotyping experiment. Despite the simplistic description given here, the pre-processing of the raw data is a major step requiring a certain level of expertise. However, most users will receive an already useable file format from a core genotyping facility. We will describe how one will receive the necessary PLINK format out of the raw data from both platforms, as this represents the actual start of the GWAS methodology in this chapter.

1. For the **Illumina® GenomeStudio** platform, creation of the necessary file format is straightforward, because a plugin, which can convert the data directly to the .ped-format, is provided by the manufacturer (*see Note 8*). Please note, in this chapter, we start from the binary file-format. Therefore, if the data is exported in ped-format from GenomeStudio, the second plink-command will need to be executed (Subheading 3.2.2). If GenomeStudio is

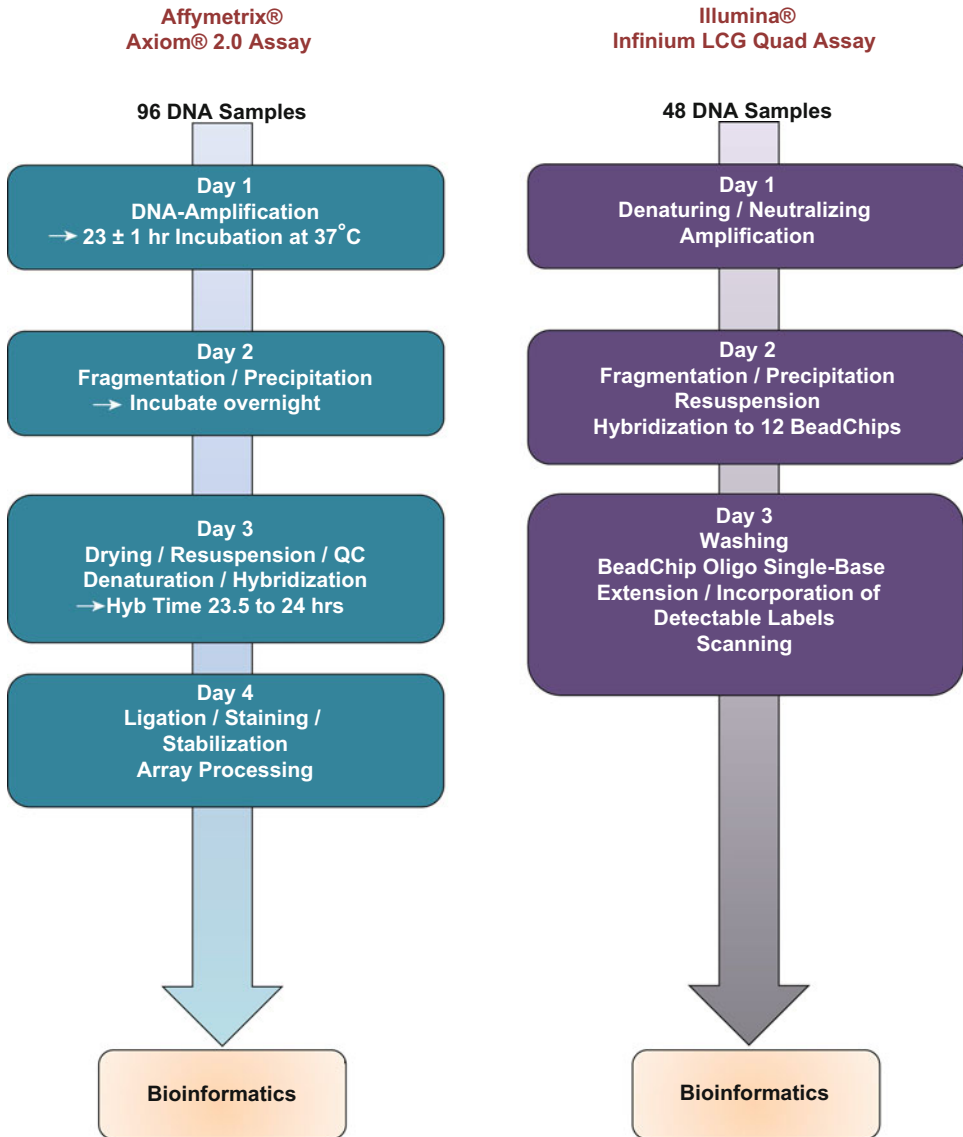


Fig. 2 Schematic protocols for the Affymetrix® and Illumina® wet lab procedures. The protocols are scalable and leave room to tighten or extend the time plan. Adopted from the manufacturer manuals

not available, and only a “final report”- and “SnpMap”-file were received, the all-in-one converting suite, “SNPConvert” can be used [18]. This software is available for download at the following github repository: https://github.com/nicolazzie/SNPConvert/tree/master/SNPConvertGUI_WinMac. Simply save the necessary .zip on the hard drive, unpack it, and run the .exe. However, because the software is written in Python 2.7, the programming language, which is available for all OS, needs to be installed. One further comment is necessary for Illumina®

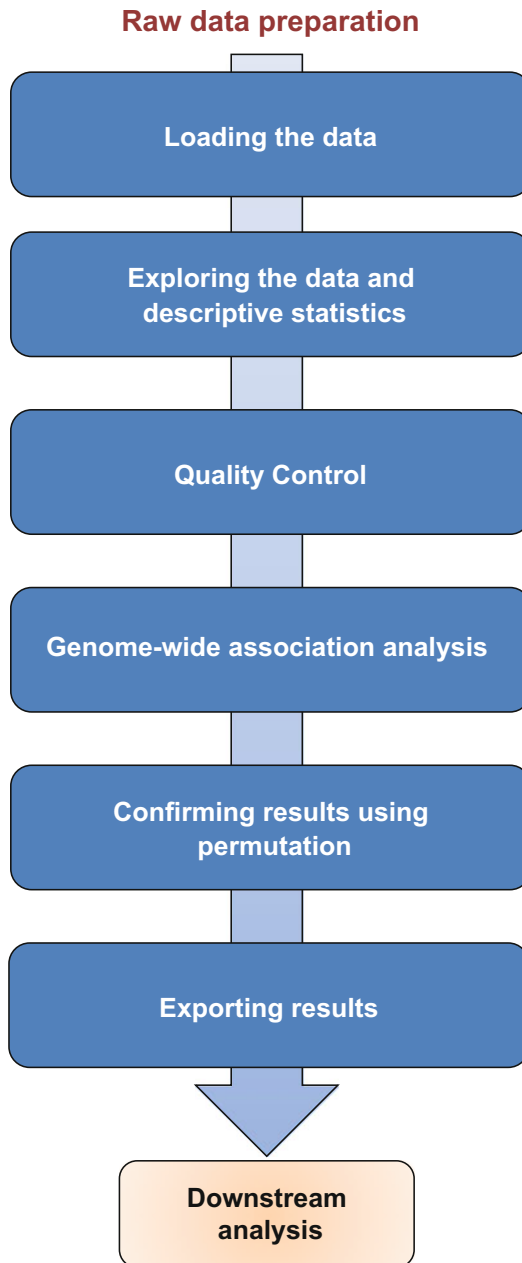


Fig. 3 Schematic workflow for the bioinformatics pipeline

arrays which contain exomic markers. Illumina[®] uses its own identifiers for markers, the so-called exm-identifier, instead of the commonly used rs-identifiers. To annotate the exm-identifiers to rs-identifiers, Illumina[®] provides auxiliary-files for different arrays at their support-homepage.

Table 3
PLINK file formats

File format	Associated file types	Function of file format	Function of the particular file
Binary	.bed .bim .fam	– Storing and processing data directly with PLINK; not human readable (encoded with 0 and 1); needs small amount of disk space	– Stores the genotypes – Mapping-file with SNP-information – Information on individuals
PED	.ped .map	– Human readable file format and therefore better to handle externally of PLINK; needs more disk space	– Stores information of one individual and genotype of all SNPs in one row – Mapping-file with SNP-information
TPED	.tped .tfam	– Transposed human readable file format; best readability to handle externally of PLINK; needs more disk space	– Stores information of one SNP and all genotypes of all individuals in one row – Information on individuals

- Exporting data from the **Axiom™ Analysis Suite** is possible in three formats: txt, PLINK (PED or TPED), and VCF. There is no limitation regarding nonhuman data. While working with this new Affymetrix® technology, one has to consider that Affymetrix® uses, similar to the exome arrays from Illumina®, its own identifiers. The structure of these is “Affx-xxx” where xxx is a number. Similar to Illumina®, Affymetrix® provides annotation-files to translate affy-identifiers to rs-identifiers. For PLINK file generation, Affymetrix® provides two methods, using either the APT command apt-format-result (<http://media.affymetrix.com/support/developer/powertools/changelog/apt-format-result.html>) or the Axiom Analysis suite. Both methods allow the user to specify the identifier to be used.
- PLINK* uses different types of data formats to achieve different goals (Table 3). All formats consist of multiple files. For further information please *see* the official PLINK homepage (Table 2).

3.2.2 Data Preparation

- For this protocol, we assume that the starting point of the analysis is the binary PLINK format.

Our test data is called: GWAS_test.bed, GWAS_test.bim, GWAS_test.fam
 Stored in folder: c:\GWAS\
 plink.exe is stored in folder: c:\PLINK\

After opening the windows cmd, navigate to the folder where the plink.exe was unzipped. To convert the unreadable binary format

to the TPED format, which will be used as input to GenABEL (the most often used approach with GenABEL), execute the following commands:

```
C:\Users\User>cd\
C:\>cd PLINK
C:\PLINK>plink --bfile c:\GWAS\GWAS_test --recode --tab --out
c:\GWAS\GWAS_test_ped
C:\PLINK>plink --file c:\GWAS\GWAS_test_ped --recode --transpose --out
c:\GWAS\GWAS_test_tped
```

2. Before the data can be explored, the files will need to be copied to Cygwin (*see* folder structure above; Subheading 2.4.2). Open Cygwin and check the files. If using Linux or Mac OS, use the Bash shell (*see* Note 9).

```
User@E00001 ~
$ cd Data/

User@E00001 ~/Data
$ head -n5 GWAS_test_tped.tped | cut -d ' ' -f1-26

1 rs1 0 10 A A G A A A A A A A G A G A A A A A A
1 rs2 0 20 C C T C C C C C C C C C T C T C C C C C
1 rs3 0 30 C T C T T T C T C T T T C T T T C T T T C T
1 rs4 0 40 T C T 0 0 C T T T C T T T T T C T C C T T T
1 rs5 0 50 C C T C C C C C C C C C C C T C C C C C T C

User@E00001 ~/Data
$ head -n5 GWAS_test_tped.tfam

ID1 ID1 0 0 1 -9
ID2 ID2 0 0 1 -9
ID3 ID3 0 0 2 -9
ID4 ID4 0 0 1 -9
ID5 ID5 0 0 2 -9
```

3. In addition to the .tped and .tfam, a phenofile, which contains phenotypic data for the cohort, is needed. This file must include the same individuals as the .tfam-file, as well as the same IDs for each individual, as defined in the .tfam-file. This file can be created with the Excel program and stored as a simple text-file (space delimited). The samples should be stored in the rows, one individual per row, and variables are stored in the columns. The number of columns is unrestricted. The phenofile should be stored within the folder where the other files are, in our case here: “c:\GWAS”. As mentioned above, this file needs to be copied to the respective Cygwin folder in order to explore it. Here we see the first four lines (including the header line) of our example phenofile (using Cygwin).

```

User@E00001 ~
$ cd Data/

User@E00001 ~/Data
$ head -n4 GWAS_test_tped.pheno.txt

IDs age sex ln_Adiponectin ln_BMI disease
ID1 74.9 1 3.063858103 3.417726684 0
ID2 68 0 NA 3.314186005 1
ID3 65.3 0 2.79971739 3.520460802 0

```

3.3 The GenABEL-Workflow

The protocol provided here is adopted from the official GenABEL tutorial (Table 2) [17]. GenABEL offers many options for different hypotheses, two of which will be described in this section. The workflow is based on first the analysis of a binary trait (case—control; Subheading 3.3.4) and second on the analysis of a quantitative trait (Subheading 3.3.4). With more complex data than that presented here, we recommend exploring the GenABEL tutorial in more detail, as it provides good examples for structured association and EIGENSTRAT analysis, as well as Mixed Models, and a combination of the latter with a structured association and subsequent meta-analysis using MetABEL. The tutorial also provides a helpful overview of what method to use to best address a specific analytical need.

From this point, we will describe all analyses using R. The first step is to start R-Studio. The text, which is written behind a “#” is a so-called code-comment. Explanations and descriptions for the command can be added, which may help to recapitulate the programming when the analysis is reassessed or the code is used for a new analysis. If the code is executed in R, this text will be ignored from the compiler.

3.3.1 Loading the Data

- Install and load the GenABEL package and convert the .tped and .tfam to GenABEL-compatible .raw-format:

```

> setwd("c:/GWAS/") # this defines the actual working directory
> install.packages("GenABEL")
> library(GenABEL)
> convert.snp.tped(tpedfile = "GWAS_test_tped.tped", tfamfile =
"GWAS_test_tped.tfam", outfile = "GWAS_test_tped.raw")

```

- Create the gwaa.data-object while reading the data into R and therefore, into the memory:

```

> GWAS_test_raw <- load.gwaa.data(phenofile =
"GWAS_test_tped_pheno.txt", genofile = "GWAS_test_tped.raw", id =
"IDs")

```

- Check the number of SNPs per chromosome:

```
> table(chromosome(GWAS_test_raw))
 1  10  11  12  13  14  15  16  17  18  19  2  20  21
22  23  3   4   5   6   7   8   9
7847 5678 5358 4903 3790 3193 2798 3033 2291 2911 1337 8215 2538 1457
1189 2113 6774 6279 6378 6472 5293 5534 4619
```

- In the .tped, the x-chromosome is defined as chromosome 23. GenABEL would identify this as an autosome, therefore the chromosome 23 must be recoded to X:

```
> GWAS_test_raw <- recodeChromosome(GWAS_test_raw,
rules=list("23"="X"))
```

- Check the chromosomes again:

```
> table(chromosome(GWAS_test_raw))
 1  10  11  12  13  14  15  16  17  18  19  2  20  21
22  3   4   5   6   7   8   9   X
7847 5678 5358 4903 3790 3193 2798 3033 2291 2911 1337 8215 2538 1457
1189 6774 6279 6378 6472 5293 5534 4619 2113
```

3.3.2 Exploring the Data

- A couple of descriptive statistics:

```
> nsnp(GWAS_test_raw) # number of snps in the analysis
[1] 100000
> nids(GWAS_test_raw) # number of samples in the analysis
[1] 100
> table(male(GWAS_test_raw)) # frequency of female and male;
                                0 = female, 1 = female
 0  1
66 34
```

- Create and explore a summary of the genotypic data (a couple of possible commands for exploration are shown); the result of the shown command is an R data frame which can be handled as such (*see Note 10*):

```
> data_summary <- summary(gtdata(GWAS_test_raw))

> data_summary[1:3,1:4] # this command only takes the first 3 lines
                        and first 4 columns of the data_summary ta-
                        ble

      Chromosome Position Strand A1 A2
rs2980300      1    825852      u  C  T
rs307378       1   1308770      u  G  T
rs3766180      1   1563420      u  T  C
```

- One can also check specific SNPs (*see Note 10*):

```
> data_summary[c("rs2980300", "rs307378"),1:4] # again for the first 4
                                                columns

      Chromosome Position Strand A1 A2
rs2980300      1    825852      u  C  T
rs307378       1   1308770      u  G  T
```

- Some other interesting measures are the SNP-call rate or the minor allele frequency:

```
> sum(data_summary$CallRate>=0.98) # prints the count of SNPs which
                                    have a callrate greater or equal
                                    to 98%

[1] 67501

> sum(data_summary$Q.2<0.05) # number of SNPs which have a minor al-
                                lele frequency smaller than 5%

[1] 10076
```

- Above we summarized the genotypic data. A summary of the sample data can also be performed. Again, the output is an R data frame:

```
> sample_summary <- perid.summary(GWAS_test_raw)

> sample_summary[1:5,c(7,8)] # shows the results for the first 5 rows
                              and the columns 7 and 8

      CallPP      Het
ID1 0.95694 0.3180555
ID2 0.96847 0.3237684
ID3 0.97639 0.3206301
ID4 0.96363 0.3228210
ID5 0.99083 0.3154224
```

“CallPP” is the sample callrate, which represents the number of SNPs successfully genotyped for a particular sample. “Het” displays the average heterozygosity of all SNPs of one sample.

- Again, the number of samples with a callrate smaller than a defined threshold (e.g., 0.95) can be checked and displayed with all summary variables:

```
> sum(sample_summary$CallPP<0.95)
> sample_summary[which(sample_summary$CallPP<0.95),]
      NoMeasured NoPoly      Hom E(Hom)      Var      F CallPP      Het
ID18      94074  93875 0.6668 0.6769 0.5167 -0.03117 0.94074 0.3331
ID50      94261  94065 0.6697 0.6774 0.5010 -0.02364 0.94261 0.3302
ID88      94717  94528 0.6607 0.6773 0.5156 -0.05170 0.94717 0.3393
ID94      94832  94637 0.6684 0.6778 0.5190 -0.02899 0.94832 0.3316
```

Unfortunately, there is no detailed description available regarding the output of “perid.summary()”. Therefore, the meaning of the output columns can only be estimated: “NoMeasured” = Number of genotyped markers per ID/sample; “NoPoly” = NA; “Hom” = average homozygosity of all SNPs of one sample; “E (Hom)” = expected value of the frequency of homozygosity; “Var” = variance; “F” = F-statistic.

3.3.3 Quality Control

Quality control is performed in three steps: (1) exclude samples and SNPs that do not fulfill specific criteria (low strictness); (2) check population for possible genetic outliers; and (3) exclude samples and SNPs that do not fulfill strict criteria.

Softened thresholds will be used in the first round, because it is possible that “check.marker()” will exclude samples from different populations or relatives, if present. It is recommended to have these samples included in the analysis when performing the second step to receive a more robust result while checking for genetic outliers. To exclude the last outliers that will not fulfill strict criteria, the quality control will be completed with a second “check.marker()” call. Thresholds such as minor allele frequency (MAF), call rate, and Hardy-Weinberg Equilibrium (HWE) are manufacturer- and SNP-chip specific.

1. Exclude samples and SNPs that do not fulfill specific criteria (low strictness)
 - In the first round, we use a SNP- and sample call rate threshold of 0.95 and a minor allele frequency cutoff of 0.01. The false discovery rate (FDR) for unacceptably high individual heterozygosity (het.fdr) and the cut-off for the HWE (p.level) are set to 0 for reasons mentioned above:

```

> qc_First_Round <- check.marker(GWAS_test_raw, callrate = 0.95,
  perid.call=0.95, p.level=0, maf=0.01 , het.fdr=0)

> summary(qc_First_Round) # produces a summary of the check.marker
  object -> not displayed

> sum(!is.na(qc_First_Round$snpok)) # number of valid SNPs after first
  qc round

[1] 83405

> sum(!is.na(qc_First_Round$idok)) # number of valid samples after
  first qc round

[1] 100

```

- We save the data-set without the outliers from the first round, and then set the remaining heterozygous X-chromosome SNPs in males to NA:

```

> GWAS_test_qc1 <- GWAS_test_raw[qc_First_Round$idok,
  qc_First_Round$snpok]

> GWAS_test_qc1 <- Xfix(GWAS_test_qc1)

no X/Y/mtDNA-errors to fix

```

2. Check population for possible genetic outliers

- First, create a genomic kinship matrix—identity by state (IBS) procedure.

```

> GWAS_test.gkin <- ibs(GWAS_test_qc1[, autosomal(GWAS_test_qc1)],
  weight="freq")

> GWAS_test.gkin[1:4, 1:4] # displays the first 4 rows and 4 columns
  of the matrix -> without colours

```

	ID1	ID2	ID3	ID4
ID1	0.494300582	7.840300e+04	7.880100e+04	7.783500e+04
ID2	-0.004456013	4.957725e-01	7.944900e+04	7.848000e+04
ID3	-0.011735060	-1.096091e-02	4.959486e-01	7.893100e+04
ID4	-0.001211697	-4.307546e-04	1.082444e-03	4.894404e-01

The numbers below the diagonal (red) are the average IBS values (orange). The numbers on the diagonal can be calculated with the equation 0.5 plus the genomic homozygosity. Above the diagonal, the amount of SNPs successfully genotyped in both individuals is displayed (white).

- Next, compute the transformation of the result to a common distance matrix, perform a classical multidimensional scaling (MDS), and plot the results as scatterplot to display possible genetic outliers. The whole GenABEL approach for the outlier detection is adapted from Price et al. [19] (*see Note 11*).

```
> GWAS_test.dist <- as.dist(0.5-GWAS_test.gkin) # transforms the
                                                kinship matrix to a
                                                distance matrix

> GWAS_test.mds <- cmdscale(GWAS_test.dist) # performs the MDS

> plot(GWAS_test.mds) # creates a scatterplot of the first two
                       principal components
```

In this example, we see one main cluster and four smaller clusters (red circles), which are potentially genetic outliers (outlier-cluster) (Fig. 4). We can find these clusters using the command “kmeans()”. “Kmeans” not only finds the outlier-clusters, but also the main sample cluster. So the argument “centers” should be the number of outlier-clusters ($n = 4$), plus the main cluster ($n = 1$)—in this case “5”. Type the following code:

```
> km <- kmeans(GWAS_test.mds, centers = 5, nstart = 1000)

> c11 <- names(which(km$cluster==1))

> c12 <- names(which(km$cluster==2))

> c13 <- names(which(km$cluster==3))

> c14 <- names(which(km$cluster==4))

> c15 <- names(which(km$cluster==5))
```

These outliers could be due to errors in sample genotyping (e.g., doubling of samples), twins or relatives in the study sample or the presence of different ethnicities. It is worth checking the phenotype data for relatives, if available, and comparing these with the result of the IBS-computation. In our example, we assume that there are four outlier-clusters.

- As all PCs in the data set contribute to the overall variance, it is worth checking other principal components, in addition to the first two computed by default using the “cmdscale” command. Compute the PCs and assess the contribution of these (“prcomp” computes all PCs, only the first six are shown):

```

> GWAS_test.pca <- prcomp(x = GWAS_test.dist, center = T)

Importance of components:
          PC1      PC2      PC3      PC4      PC5      PC6
Standard deviation  0.10383 0.10123 0.09628 0.08703 0.07833 0.07604
Proportion of Variance 0.03934 0.03740 0.03383 0.02764 0.02239 0.02110
Cumulative Proportion 0.03934 0.07674 0.11057 0.13821 0.16060 0.18170

```

As shown, the first six PCs explain only 18% of the variance, suggesting that other PCs contribute to the variance in a not negligible way. Therefore, it is worthwhile to check at least these PCs. We wrote a short R-function “plot.GeneticOutlier()” for doing so, which can be downloaded from: https://github.com/TobiWo/Plotting_GeneticOutliers (description, installation and further comments see repository). The function displays multiple MDS-plots and automatically marks the clusters/outliers from the first MDS-plot (PC1 vs. PC2), in all other MDS-plots. Further, if the argument “return.main” is set to true, it returns the sample-IDs, which should be kept in the analysis, i.e., the main-sample-cluster (see source code for detailed description of all arguments). If the above samples were real outliers, they would also build clusters in the other MDS-plots. To compute more than two PCs, the

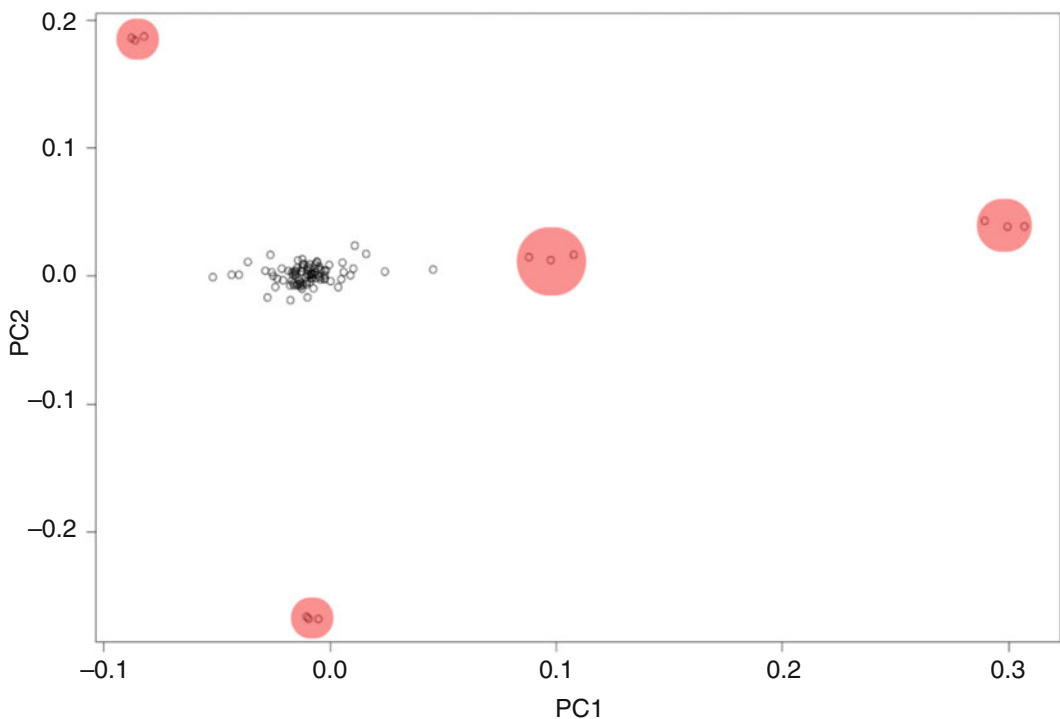


Fig. 4 Scatterplot displaying the results of the multidimensional scaling (MDS) analysis to display possible genetic outliers. The first two principal components are shown. One main cluster and four smaller clusters (*red circles*) can be found

`cmdscale()` call must be performed again while specifying the number of desired principal components. In the following example, only the first four will be plotted. Compute the following:

```
> GWAS_test.mds <- cmdscale(GWAS_test.dist, k = 6) # k specifies the
                                                    number of PCs
                                                    calculated

> plot.GeneticOutlier(x = GWAS_test.mds, clusters = 5, n.pc = 4)

Merging plots...
NULL
```

As shown in Fig. 5, in the first MDS-plot, one small cluster (blue circle) is not an outlier. Instead, three other samples on the left of the main cluster are marked as outliers (red circle). One would not identify these while looking only at the first MDS plot. However, checking the other plots reveals that indeed, this cluster (red circles) appears to be a real outlier-cluster (Fig. 5). To be sure

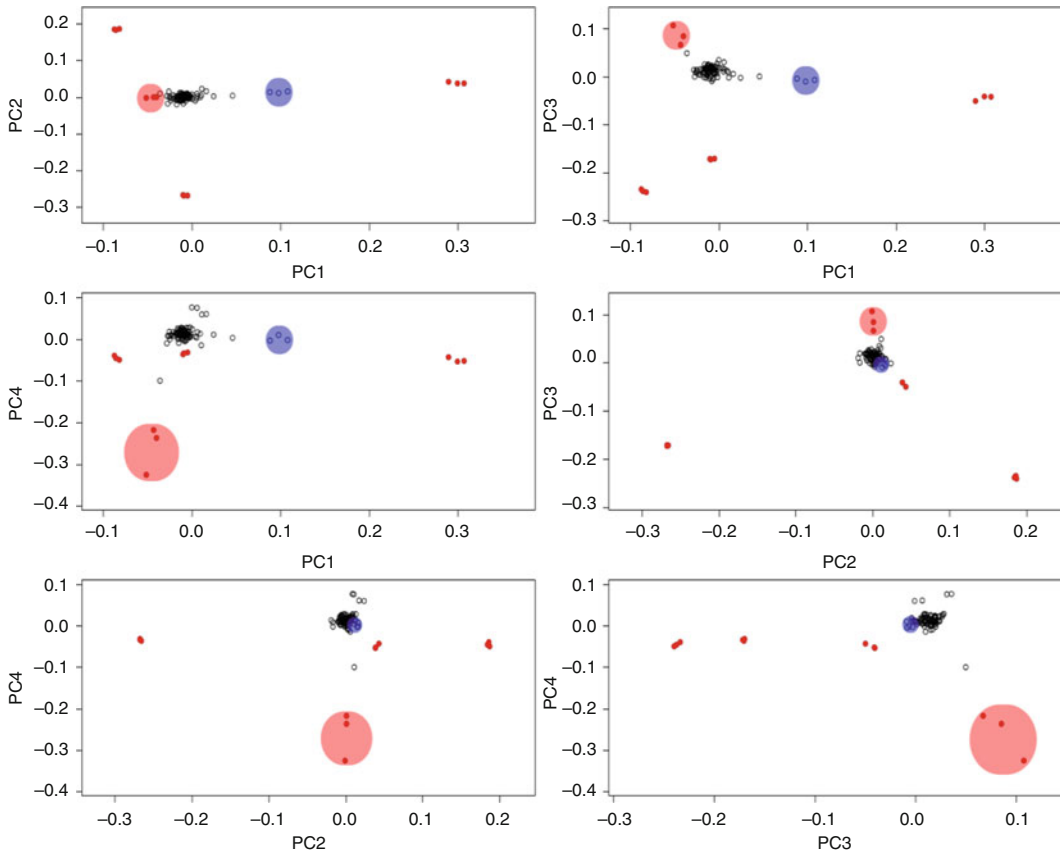


Fig. 5 Scatterplots displaying the multidimensional scaling (MDS) analysis to identify samples outlier clusters. Principal components one to four are shown. The small cluster (*blue circle*) in the MDS-plot is not an outlier. Instead, three other samples on the left of the main cluster are marked as outliers (*red circle*)

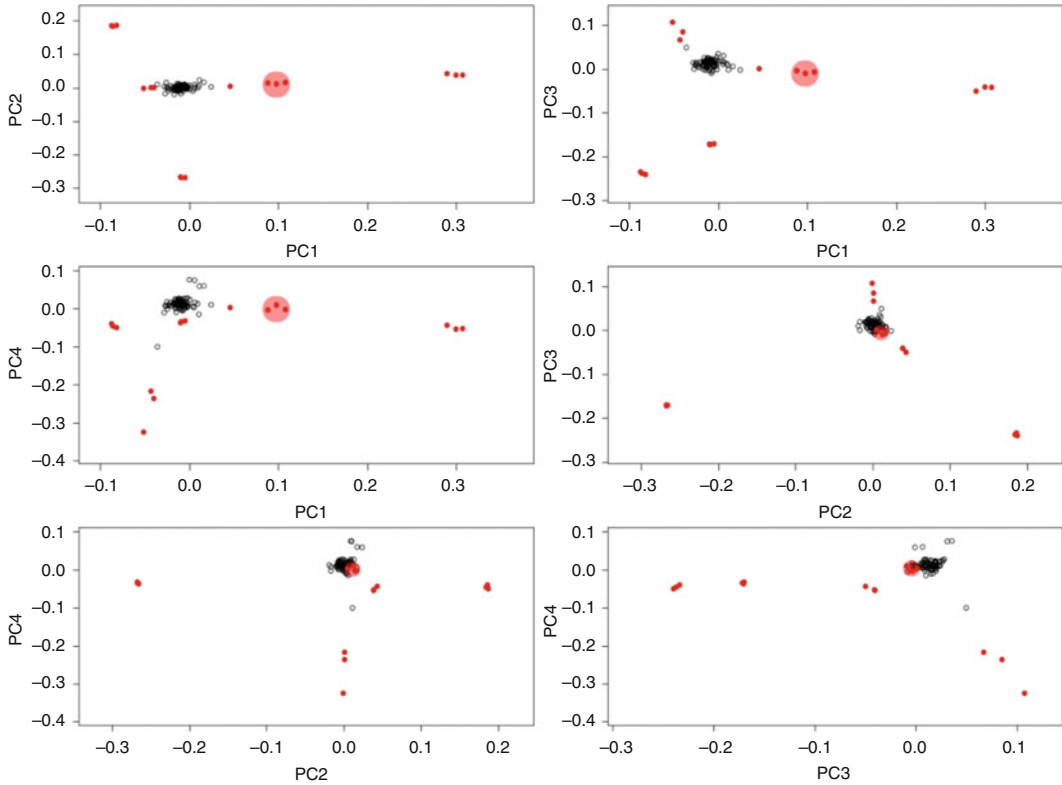


Fig. 6 Scatterplot displaying the multidimensional scaling (MDS) analysis for five outlier clusters. For the first three plots, the additional outlier-cluster (*red circles*) is true, but for the remaining three, it is not

about the unclear samples (blue circles), another check for five outlier-clusters (+1 main cluster = 6) will be performed (Fig. 6). Additionally, we return the sample IDs of the main cluster, which we will keep in the analysis:

```
> IDs_to_keep <- plot.GeneticOutlier(x = GWAS_test.mds, clusters = 6,
  n.pc = 4, return.main = TRUE)
```

```
Merging plots...
```

For the first three plots, the additional outlier-cluster (red circles) is true, but for the remaining three, it is not (Fig. 6). We also checked other PCs with our function (data not shown) and could see that these three samples form an outlier-cluster in most of the plots. Therefore, we decided to exclude these samples. However, this is a case-to-case decision that deserves careful consideration.

3. Exclude samples and SNPs that do not fulfill strict criteria

We clean the data while supplying the IDs of the main-cluster to the `gwaa.data`-object. Then we proceed with the second “`check.marker()`” call on the remaining data-set. Instead of `p.level = 0`, we now use the manufacturer-defined threshold for HWE. The `het.fdr`-argument is not displayed because we used the default value of 0.01:

```
> GWAS_test_qc1_NoGeneticOutlier <- GWAS_test_qc1[IDs_to_keep,]
> qc_Second_Round <- check.marker(GWAS_test_qc1_NoGeneticOutlier,
  callrate = 0.95, perid.call=0.95, p.level=0.0001,
  maf=0.01)
> summary(qc_Second_Round)
$`Per--SNP fails statistics`
      NoCall NoMAF NoHWE Redundant Xsnpfail
NoCall      2124    4    4         0         0
NoMAF        NA   209    0         0         0
NoHWE        NA   NA   22         0         0
Redundant    NA   NA   NA         0         0
Xsnpfail     NA   NA   NA         NA         0

$`Per--person fails statistics`
      IDnoCall HetFail IBSFail isfemale ismale isXXY otherSexErr
IDnoCall      0      0      0         0         0      0         0
HetFail       NA      0      0         0         0      0         0
IBSFail       NA     NA      0         0         0      0         0
isfemale     NA     NA     NA         0         0      0         0
ismale       NA     NA     NA         NA         0      0         0
isXXY        NA     NA     NA         NA         NA      0         0
otherSexErr   NA     NA     NA         NA         NA     NA         0
```

As shown, 2363 SNPs, but no additional samples, were dropped in the second qc-round. We excluded these SNPs, creating a final data-set, which will be used for the GWAS:

```
> GWAS_test_clean <- GWAS_test_qc1_NoOutlier[qc_Second_Round$idok,
  qc_Second_Round$snpok]
```

3.3.4 Genome-Wide Association Analysis

For the purposes of illustration, we will perform two different GWAS on the sample data: a qualitative (binary/case-control) analysis and a quantitative trait analysis. First, we check the phenotype data:

```
> descriptives.trait(GWAS_test_clean)
```

	No	Mean	SD
id	86	NA	NA
age	86	63.869	9.488
sex	86	0.349	0.479
ln_Adiponectin	82	2.785	0.383
ln_BMI	85	3.385	0.183
disease	86	0.372	0.486

1. Case-control analysis

The qualitative case-control analysis will be conducted on the disease phenotype, which is simply defined as “yes” (equals 1) or no (equals 0). The “descriptives.trait()” calculates a mean and a SD also for the binary traits sex and disease, but these results should be ignored.

The association analysis itself is fairly easy to perform. GenABEL uses the same formula-context as most statistical functions in R. The variable before the “~” is the response variable, after the “~”, the predictor variables are added. Multiple covariates are connected with a plus sign. The markers for the actual GWAS are not specifically included in the formula, as this is done by GenABEL internally. In our example, we adjust for age and sex, as these variables are known to influence the association between the markers and our hypothetical disease phenotype. Let us conduct the GWAS as follows:

```
> GWAS_lreg_bin <- mlreg(formula = disease~sex+age, data =
  GWAS_test_clean, trait.type = "binomial")

> lambda(GWAS_lreg_bin) # displays the genomic inflation factor which
  should be: 1<=λ<1.05
  # λ<1 are forced to be 1
  # this result could be due to the low sample
  size in our disease--groups but it is not a
  problem at all > no inflation present

$estimate
[1] 0.98898

$se
[1] 0.0001409786
```

Following the analysis, the results are saved and the false discovery rate (FDR)-corrected p-values are added to the data-frame. The top ten most significantly associated markers are displayed (*see Note 12*):

```

> GWAS_result_bin <- descriptives.scan(data = GWAS_lreg_bin, top =
  nsnp(GWAS_test_clean))

> GWAS_result_bin <- GWAS_result_bin[,1:11] # deletes the last four col-
  umns of the result data-
  frame, since these have
  only NAs under mlreg call

> GWAS_result_bin$p.FDR <- p.adjust(p = GWAS_result_bin$Pcldf, method =
  "BH")

> GWAS_result_bin[1:10,]

```

	CHR	A1	A2	N	effB	se_effB	Pcldf	p.FDR
rs1530790	16	A	G	85	2.047412	0.5230054	6.400488e-05	0.9273647
rs12458379	18	A	G	86	2.153743	0.5607248	8.771860e-05	0.9273647
rs2163339	7	G	T	86	1.581759	0.4203796	1.218828e-04	0.9273647
rs1012264	2	A	G	86	2.246082	0.6029936	1.425507e-04	0.9273647
rs7320337	13	T	C	83	-1.775926	0.4870054	1.962582e-04	0.9273647
rs322679	3	T	G	86	1.637187	0.4514762	2.130416e-04	0.9273647
rs979775	6	G	A	86	2.270543	0.6261541	2.131496e-04	0.9273647
rs4259369	7	T	C	83	1.538657	0.4258574	2.246679e-04	0.9273647
rs17164903	5	C	G	86	-2.508894	0.6953466	2.291840e-04	0.9273647
rs11175790	12	G	A	85	1.644059	0.4697830	3.520228e-04	0.9273647

As shown from the top10 output (column “Pcldf”), none of the SNPs in this example reached the commonly used genome-wide significance level of $p \leq 1 \times 10^{-8}$ which is based on the conservative multiple testing procedure of Bonferroni [20]. The Bonferroni procedure indicates that for testing of n independent hypotheses on one dataset, the statistical significance for every single hypothesis is $1/n$ of the significance which would be used for testing only one hypothesis. In addition to Bonferroni-corrected p-values, false discovery rate (FDR) corrected p-values are calculated (column p.FDR). In GWAS, a FDR-threshold of 5–20% is commonly used [21]. In our analysis, even with a threshold of 20%, we found no statistically significant evidence for association between SNPs and phenotype.

Another interesting output is the column “effB” which, in case of a binary trait, represents the odds ratio (OR) on a log-scale. One can calculate the OR and corresponding confidence interval for the displayed top hit by exponentiation of the values from the columns “effB” and “se_effB” (latter is standard error):

```

> exp(GWAS_result_bin$effB[1]) # Odds ratio

[1] 7.747824

> exp(GWAS_result_bin$effB[1]-
(1.96*GWAS_result_bin$se_effB[1])) # lower Confidence interval

[1] 2.77964

> exp(GWAS_result_bin$effB[1]+
(1.96*GWAS_result_bin$se_effB[1])) # upper Confidence interval

[1] 21.59588

```

This result can be interpreted as a 7.7-fold higher risk developing the disease when one effect allele is added to the genotype. However, because there is no evidence of statistical significance, we would not consider the SNP as a real hit.

2. Quantitative analysis

In the following example, the quantitative analysis will be performed using plasma-levels of adiponectin, a protein involved with metabolic and hormonal processes. Because the original data points for adiponectin were not normally distributed, we would expect that the residuals of the actual models would not be normally distributed, which violates a major assumption for linear regression. Therefore, the data points need to be *ln*-transformed prior to analysis (*see* output of “descriptives.trait()”).

We conduct the same function as for the binary trait example described above, and also adjust for sex and age. However, because we are now evaluating a quantitative trait, the argument “trait.type” must be changed from “binominal” to “gaussian”:

```

> GWAS_lreg <- mlreg(formula = ln_Adiponectin~sex+age, data =
  GWAS_test_clean, trait.type = "gaussian")

> lambda(GWAS_lreg) # displays the genomic inflation factor which
  should be: 1<=λ<1.05

$estimate
[1] 1.03744

$se
[1] 0.0001409786

```

Again we save the results, add false discovery rate (FDR)-corrected p-values, and display the top ten SNPs (*see* **Note 12**):


```

> GWAS_result_quant <- descriptives.scan(data = GWAS_lreg, top =
  nsnp(GWAS_test_clean))

> GWAS_result_quant <- GWAS_result_quant[,1:11] # deletes the last four
  columns of the result
  data--frame, since
  these have only NAs
  under mlreg call

> GWAS_result_quant$p.FDR <- p.adjust(p = GWAS_result_quant$Pcdf, meth-
  od = "BH")

> GWAS_result_quant[1:10,]

```

	CHR	A1	A2	N	effB	se_effB	Pcdf	p.FDR
rs7003979	8	C	A	82	-0.2875540	0.05866375	1.490757e-06	0.07391030
rs1955392	16	T	A	79	-0.2216730	0.04560690	1.824022e-06	0.07391030
rs4852037	2	A	G	82	-0.6733711	0.14130660	2.889218e-06	0.07804838
rs785495	1	T	C	79	0.2443125	0.05301194	6.047893e-06	0.12253182
rs10116672	9	T	A	82	-1.0936462	0.24189379	9.044015e-06	0.14658720
rs16947096	16	C	T	82	-0.2244107	0.05051180	1.289763e-05	0.17420620
rs7184271	16	T	C	80	-0.2205638	0.05033758	1.693393e-05	0.19025543
rs991486	X	T	C	78	-0.5846680	0.13450663	1.975832e-05	0.19025543
rs598130	1	G	A	81	-0.2998850	0.06972320	2.413327e-05	0.19025543
rs3860052	12	G	T	81	-0.2755493	0.06426558	2.558395e-05	0.19025543

As with the binary analysis, none of the SNPs reach genome-wide significance levels of association. However, if we set the FDR-threshold to 20%, all of the SNPs show statistically significant evidence for association with adiponectin levels, yet, we would expect that two of the ten are false positive hits. The “effB” column represents the actual beta or coefficient estimate of a linear regression. Because we used *ln*-transformation to achieve a normal distribution of adiponectin values, these results are not easily interpretable. If the transformation had not been applied, we would interpret “effB” as the increase or decrease of adiponectin (in $\mu\text{g}/\text{ml}$), while adding one allele “A2” to the genotype.

3.3.5 Permutation

In addition to Bonferroni or FDR-corrections, permutation analysis is another approach to address the problem of multiple testing [22]. Like these methods, the general purpose of a permutation analysis is to assess whether the GWAS result appeared by chance or is statistically meaningful. Permutations are helpful to determine candidate SNPs for pathway analyses. A recent study showed that permutations are more suitable than simple multiple testing correction methods for detecting false positive hits [23]. There are different permutation methodologies available [23], and GenABEL itself offers a permutation option within the function “qtscore()”. This statistical test works slightly different to “mlreg()”, which we conducted for the first GWAS. This results in a minimal deviation

regarding the output of the function. If one decides to use the permutation functionality of “`qtscore()`”, it must also be performed for the original GWAS (Subheading 3.3.4) to maintain consistency over the different analyses. In accord, we repeated the initial GWAS, checking then our results while using 500 permutations with “`qtscore()`” (see **Note 13**). Here we continue with the quantitative trait, but the procedure would be the same for the binary analysis, except the argument “`trait.type`” would be changed to “`binomial`”:

```
> GWAS_qts <- qtscore(formula = ln_Adiponectin~sex+age, data =
  GWAS_test_clean, trait.type = "gaussian")

> GWAS_score_permutation <- qtscore(formula = ln_Adiponectin~sex+age,
  data = GWAS_test_clean, trait.type = "gaussian", times = 500)

> GWAS_result_permutation <- descriptives.scan(data =
  GWAS_score_permutation, top = nsnp(GWAS_test_clean))

> GWAS_result_permutation[1:10,]

```

	CHR	A1	A2	N	effB	se_effB	Pc1df
rs7003979	8	C	A	82	-0.2874777	0.06542787	0.476
rs4852037	2	A	G	82	-0.6660878	0.15574948	0.654
rs1955392	16	T	A	79	-0.2117746	0.04988861	0.712
rs10116672	9	T	A	82	-1.0763731	0.26293538	0.932
rs785495	1	T	C	79	0.2286694	0.05672256	0.974
rs16947096	16	C	T	82	-0.2152575	0.05401008	0.982
rs991486	X	T	C	78	-0.5376237	0.13492631	0.982
rs3860052	12	G	T	81	-0.2740066	0.06953335	0.990
rs7184271	16	T	C	80	-0.2134367	0.05410649	0.990
rs598130	1	G	A	81	-0.2936361	0.07490558	0.994

While conducting the permutation analysis, the significant threshold is now 0.05 and defined as:

$$\frac{n_{\text{array.p} < \text{o.p}}}{n_{\text{permutations}}}$$

“array.p” represents a collection of the minimal p-values obtained in each permutation analysis and “o.p” is the p-value for a particular SNP obtained in the initial analysis. In our GWAS, the permutation reveals that none of the top hits reaches the significant threshold, indicating that the original results occurred by chance. Independent of such a result, it is always worth checking databases for the biological background of your top hits. If one of the markers lies within a gene that has biological relevance, which can be confirmed experimentally, then the p-values are not that important. However, a GWAS is a hypothesis-free approach to identify possible candidates for further analysis, and should therefore produce trustworthy results.

3.3.6 Post Analysis

After performing the association tests, it is necessary to check whether the results show signs of confounding, for example, from differences in population structure. Both quantile-quantile- (Q-Q) plots and Manhattan plots are common tools for identifying data confounding. To demonstrate this, we will continue with the results from the quantitative association analysis, referring to the output from “mlreg()”; however, the approach would be the same for binary trait analysis. To plot the data, we use the R-package “qqman” (<https://github.com/stephenturner/qqman>) [24]. Because the functions of “qqman” need a special input-data format, we must first edit the original results (dataframe “GWAS_result_quant” in Subheading 3.3.4):

```
> GWAS_result_plot <- GWAS_result_quant[,c(1,2,11)] # saving columns
                                                    1,2 and 11

> GWAS_result_plot$SNP <- rownames(GWAS_result_plot)#creating new
                                                    column SNP with
                                                    rownames of
                                                    GWAS_result_plot

> GWAS_result_plot <- GWAS_result_plot[,c(4,1:3)] #change order of
                                                    columns

> colnames(GWAS_result_plot)[2:4] <- c("CHR", "BP", "P") #change column
                                                    names

> levels(GWAS_result_plot$CHR)[levels(GWAS_result_plot$CHR)=="X"] <-
"23" #change x-chromosome to 23 since we need numeric annotation for
plotting

> GWAS_result_plot$CHR <- as.numeric(as.character(GWAS_result_plot$CHR))
# change data-type of column CHR to numeric
```

Next, “qqman” is installed and loaded and a Q-Q-plot is made (Fig. 7):

```
> install.packages("qqman")

> library(qqman)

> qq(pvector = GWAS_result_plot$P) # here we refer to the p-value column
of GWAS_result_plot
```

The Q-Q-plot in Fig. 7 displays the expected distribution of p-values compared to the observed p-values. All black dots should be on the red line, except those on the upper right since these refer

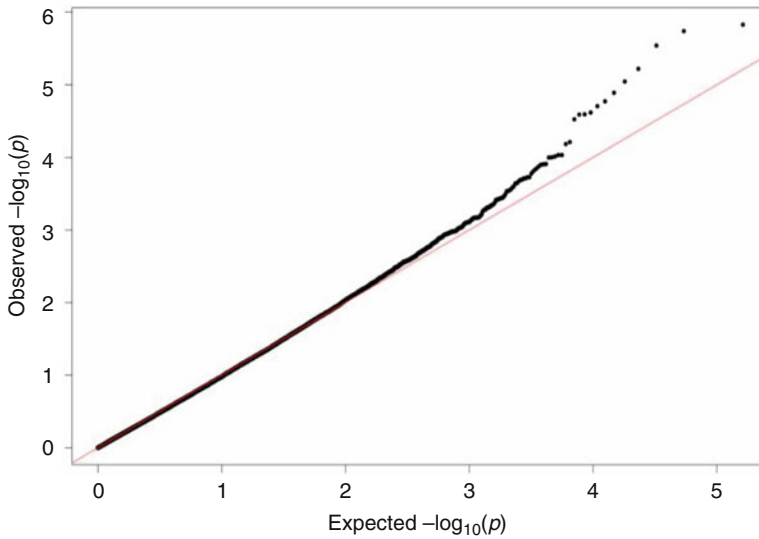


Fig. 7 Quantile-quantile (Q-Q) plot. Displays the expected distribution of p-values compared to the observed p-values

to the probable true associations. From this analysis, we can conclude that our data showed no confounding.

For the Manhattan plot, we run the following command, where *x* is the data being referred to:

```
> manhattan(x = GWAS_result_plot, col=c("black", "cadetblue1", "green",
    "red", "blue", "pink", "gray47", "gold"), genomewideline =
    F, logp = T, ylim=c(1, 8))
```

The function is designed to search for a column marked as “P” within the input-data. The command also contains features to change the color (argument “col”) and set the y-axis limitations (argument “ylim”). Because we are excluding p-values over a certain threshold, this will reduce the workload for the machine. Here we set $y = 1$ as minimum and $y = 8$ as maximum. This excludes all p-values larger 0.1 (Fig. 8).

In the Manhattan plot (Fig. 8), the blue line represents the suggestive significance threshold. While there is no defined threshold for suggestive significance, $p < 10^{-5}$ is commonly used. Duggal et al. [25] determined that different genotyping technologies or databases lead to different significant thresholds for the resulting GWAS associations. Beside the strength of the association, the Manhattan plot can also reveal problems in the data. If points were scattered or the top hits were single points, rather than clusters of associated markers (so-called chimneys), the data should be reevaluated. While the sample data initially gave the impression of isolated top hits, closer examination reveals that SNPs on

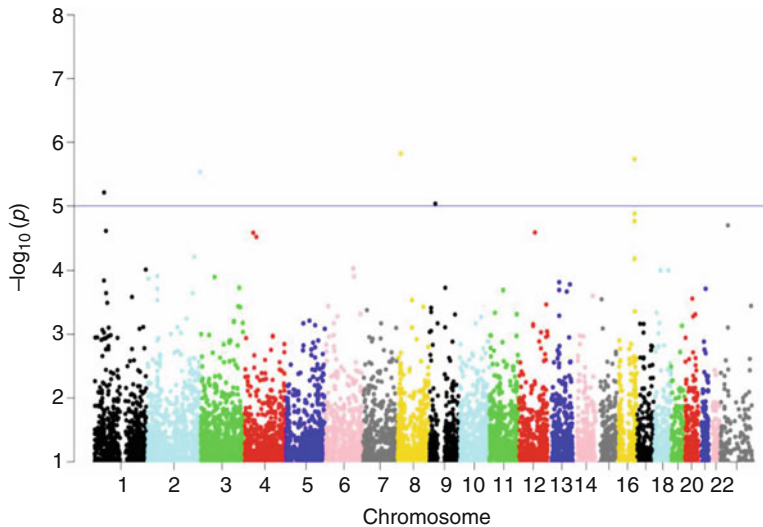


Fig. 8 Manhattan Plot. The *blue line* represents the suggestive significance threshold

chromosomes 1 and 16 are flanked by other markers. However, SNPs on other chromosomes do appear to be single markers and this could be due to the fact that we extracted the data randomly from a larger dataset, which may have lacked flanking SNPs.

3.3.7 Imputation

If genotyping is performed on the Illumina[®] or Affymetrix[®] platform, the number of SNPs interrogated represents only a fraction of those within the human genome. Although these SNPs were chosen for specific reasons (e.g., tagging SNPs with high LD to a maximum of other SNPs), the potential for missing interesting markers remains a possibility. The method of imputation offers a way to add missing data points into the dataset [26]. Imputation in genetics is based on known haplotypes of a population. In this method, algorithms infer genotypes for the missing SNPs based on a reference panel. In the past, inference used genotypes from the HapMap database, but now, the *1000 Genomes Project* database (<http://www.internationalgenome.org/>) is mostly used, as it offers information on millions of markers, and makes it possible to increase imputation quality of low-frequency variants, which scale with the overall size of the reference panel [27].

Imputation requires several checks for quality control (pre and post), as one might be faced with mapping problems and it is important to know how to handle imputed data. The above-mentioned algorithms are packed in freely available software making imputation feasible for even novice researchers. A recent paper for the “Beagle” software compared the most widely used packages over a variety of parameters and presented a comprehensive theoretical background for imputation [28]. Although, “Beagle” in the current version is memory efficient and very fast,

computer power, in general, must be considered as a limitation for imputation. Therefore, we recommend a small server-cluster, with at least 8–12 CPU cores and 64–128 GB of memory when imputing from the *1000 Genomes Project* database. This is even more pertinent when conducting a GWAS based on imputed data. Here, GenABEL, or R in general, will come to its limitations quickly, as it is not a very fast programming language compared to C/C++, for example. However, for imputed data, ProbABEL, which belongs to the GenABEL suite, was designed [29].

3.3.8 Analysis

Conclusion

From our analyses, we can conclude that the presented dataset is valid and shows no signs of confounding or related problems. The binomial analysis found no evidence for statistically significant association of any marker with the disease phenotype. Likewise, the quantitative analysis revealed no evidence for genome-wide statistical significance, although we found some suggestive hits. Multiple testing methods produced an unclear picture: while the FDR-procedure suggested the top ten hits were significant, permutation did not. As mentioned, in the face of biological relevance, biological experimentation may be indicated, as well as appropriate data analysis (*see* Subheading 3.4), or at the very least, validation of findings in a replication cohort. The latter is especially important for confirming associations in independent cohorts. Because validation is necessary to increase the explanatory power of a GWAS, one should always be conducted.

3.4 Further Downstream Analysis

3.4.1 SNP Annotation & Resources for Data Interpretation

The top hits from any GWAS need to be annotated, interpreted, and set within a biological context. Further, not only the lead SNP should garner attention, but also the whole LD cluster. We recommend checking whether associated SNPs cause changes in the amino acid sequence, disrupt predicted transcription factor binding sites, or have already been implicated in other diseases. A reasonable first step to determine whether the signal has already been detected for the trait of interest, or any other trait, is using the GWAS catalog [4, 5] (Table 4). A freely available and widely used tool for annotation is ANNOVAR, which can complete these steps [30]; however, because it is a command-line based program, it is not as easy to use as a general windows program. It is written in the programming language Perl. There is a web-based version of ANNOVAR, wANNOVAR, which provides good functionality, although it is not as complete as ANNOVAR [31, 32]. Alternatively, the Variant Effect Predictor (VEP) from Ensembl [33] can be used, either as a stand-alone Perl script or a web interface, which is very user-friendly and does not require command-line experience (Table 4). Other packages include SnpEFF [34], Exomiser [35], and VarioWatch [36]. We recommend checking the functionalities of these different tools to decide which the best approach for the analysis in question is. Furthermore, there are many excellent databases supporting

Table 4
Resources for data interpretation

Resource	Link	Description
<i>Databases</i>		
GWAS catalog	https://www.ebi.ac.uk/gwas/	Collection of all published GWAS assaying ≥ 100.000 SNPs and all SNP-trait associations
NCBI	https://www.ncbi.nlm.nih.gov/	National Center for biotechnology information
UCSC genome browser	http://genome.ucsc.edu/	Graphical viewer for aligned genome annotations
Ensembl	http://www.ensembl.org/index.html	Genome browser for vertebrate genomes: Gene annotation, alignments, predictions, regulatory function and collection of disease data.
<i>Software & Web-interfaces</i>		
ANNOVAR	http://annovar.openbioinformatics.org/en/latest/	Functional annotation of genetic variants
wANNOVAR	http://wannovar.wglab.org/	Web-based access to most functionalities of the ANNOVAR software for SNP annotation
Variant effect predictor (VEP) (Ensembl)	http://www.ensembl.org/info/docs/tools/vep/index.html	Determines the effect of your variants on genes, transcripts, protein sequence, as well as regulatory regions
wVEP	http://www.ensembl.org/Tools/VEP	Web interface for VEP
Snpeff	http://snpeff.sourceforge.net/	Genetic variant annotation and effect prediction toolbox
Exomiser	http://www.sanger.ac.uk/science/tools/exomiser	Annotation, filtering and prioritizing likely causative variants according to user-defined criteria
VarioWatch	http://genepipe.ncgm.sinica.edu.tw/variowatch/main.do	Annotation on human genomic variants

The resources given here are examples and the table does not claim to be complete

literature searches (e.g., NCBI) and genome browsing (e.g., UCSC Genome Browser [37] and Ensembl [38]). Certainly, there are more tools available than the ones presented here and many of them share the same objective. The omics tools website (<https://omicstools.com/gwas-category>) lists many available programs for annotation.

3.4.2 Expression Quantitative Trait Locus (eQTL) Studies

When gene expression data are available, genotype data can be used to identify eQTLs. There are a number of software packages available to do eQTL analysis, including “Matrix eQTL”, which is an

R-package [39]. For an R extension, “Matrix eQTL” is extremely fast, because the authors used special matrix-based methods for the algorithm calculations. Combined with the fact that it is a user-friendly software with a good tutorial, “Matrix eQTL” is a good choice for this kind of analysis.

3.4.3 Phylogenetic Module Complexity Analysis—PMCA

Identifying the disease-causing or trait-modulating variant is the main challenge following GWAS, because association signals usually tag large LD groups, and the majority of common genetic variants are located in noncoding regions. PMCA is a computer-aided process developed to address this issue. The PMCA method leverages conserved co-occurring transcription factor binding site (TFBS) patterns within *cis*-regulatory modules to predict *cis*-regulatory variants, i.e., variants affecting gene expression [40]. This method is based on the assumption that important DNA sequences have persisted throughout evolution across different species and that variants modulating gene regulation are major contributors to common disease risk.

3.4.4 Gene-Based Genome-Wide Association Study

Here we directly refer to a program developed at the University of Hong Kong by Li et al. [41, 42] called “Knowledge-based mining system for Genome-wide Genetic studies” or KGG. This program combines several powerful tests such as gene-based or gene-pair interaction-based associations. The input for KGG is simply the list of p-values received from the GWAS. KGG connects to several databases and performs a comprehensive secondary analysis of the original GWAS result. Because it is based on a graphical user interface and has a very good user manual, the workflow is straightforward.

3.4.5 Analyses with Single Markers

Once the SNP-chip data are available, new hypotheses regarding specific SNPs present in the data can be developed, for example, whether a particular marker of interest is associated with a specific phenotype. In such a case, single SNPs can be extracted from the data, recoded to the genetic model of interest, and tested for association. For someone not familiar with programming, this is not an easy task. However, we wrote a small program, called “gwasrecode”, to facilitate single SNP extraction (<https://github.com/TobiWo/gwasrecode>). The program is based on the .tped-format, which is also the basis for the workflow described above. An interesting feature of the program is that it can also extract markers from exome data-sets generated by an Illumina platform, which use special exm-identifiers for markers instead of the more common rs-identifiers. While the program is in an early stage of development, its overall functionality has been successfully tested in different scenarios.

4 Notes

1. Recommended reading for GWAS: Bush WS, Moore JH (2012) Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol.* 8(12):e1002822.
2. Recommend reading for statistical significance testing and power calculation:
Sham PC, Purcell SM (2014) Statistical power and significance testing in large-scale genetic studies. *Nat Rev. Genet.* 15:335–346. and McCarthy MI et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet.* 9:356–369.
3. Further online resources/resources for download for statistical power and sample size calculations (selection): http://www.statisticalsolutions.net/pssZtest_calc.php; <http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize>; <http://biomath.info/power/>
4. Affymetrix[®]–Axiom[®] and Illumina[®] offer arrays with different design, coverage, and price, as well as custom solutions. Therefore, manuals may differ depending on the particular array. Please see the manuals on the manufacturers' homepages.
5. For analysis procedures, all samples should be processed using the same platform. In case different genotyping platforms were used, bioinformatics adaptations are necessary to analyze the samples in one batch.
6. We have not yet used GWASTools, and cannot give a conclusion about the overall program.
7. DNA quality: a 260/280 ratio of ~1.8 is generally accepted as “pure” for DNA; expected 260/230 values are in the range of 2.0–2.2. DNA cleanup can be performed using a variety of methods.
8. http://support.illumina.com/array/array_software/genomestudio/downloads.html.
9. More information on both file formats, .tfam and .tped, can be obtained here: <https://www.cog-genomics.org/plink2/formats#tfam>.
10. The “u” in the “Strand” column means that there is no strand information available for the data. Therefore, it is undefined. This information is not necessary to conduct a GWAS; however, it can be useful to know the strand, especially for the top hits. An easy way to get this information is the Biomart-project by Ensembl (<http://www.ensembl.org/biomart/martview/3bfe0120c94ba7be65e52ce273af489b>). Here you can browse through different data sets and filter for particular SNPs or even specific SNP-chips.

11. The following steps will take several minutes, depending on the size of the data set and the speed of your machine.
12. You will notice that your output looks different to what is displayed here. We cut the data-frame for space considerations and only show the most important information. You will recognize a column called “P1df”. These are the p-values before genomic control adjustment. “Pc1df” are then the p-values after the adjustment and the ones you should refer to.
13. Depending on the size of your data, this can take some time. Again, we cut the result-data-frame to display only the most important information.

Acknowledgments

We would like to cordially thank Peter Kovacs, head of the research group Genetics of Obesity and Diabetes, and our colleagues for their everlasting scientific and personal support.

Funding: Tobias Wohland is funded by the IFB AdiposityDiseases (AD2-6E95). Dorit Schleinitz is funded by the Boehringer Ingelheim Foundation and by a Collaborative Research Center (C1, CRC1052).

References

1. LaFraniere T (2009) Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res* 37:4181–4193
2. Bush WS, Moore JH (2012) Chapter 11: genome-wide association studies. *PLoS Comput Biol* 8:e1002822
3. Kemper KE, Deatwyler HD, Visscher PM, Goddard ME (2012) Comparing linkage and association analyses in sheep points to a better way of doing GWAS. *Genet Res Camb* 94:191–203
4. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorf L, Parkinson H (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42 (Database issue):D1001–D1006
5. Burdett T (EBI), Hall PN (NHGRI), Hastings E (EBI), Hindorf LA (NHGRI), Junkins HA (NHGRI), Klemm AK (NHGRI), MacArthur J (EBI), Manolio TA (NHGRI), Morales J (EBI), Parkinson H (EBI) and Welter D (EBI). The NHGRI-EBI Catalog of published genome-wide association studies. Available at: www.ebi.ac.uk/gwas. Accessed November 2016
6. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gómez-Marín C, Aneas I, Credidio FL, Sobreira DR, Wasserman NF, Lee JH, Puviondran V, Tam D, Shen M, Son JE, Vakili NA, Sung HK, Naranjo S, Acemel RD, Manzanares M, Nagy A, Cox NJ, Hui CC, Gomez-Skarmeta JL, Nóbrega MA (2014) Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 507:371–375
7. Habek M, Brinar VV, Borovecki F (2010) Genes associated with multiple sclerosis: 15 and counting. *Expert Rev Mol Diagn* 10:857–861
8. Sham PC, Purcell SM (2014) Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* 15:335–346
9. McCarthy MI et al (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9:356–369
10. Faul F, Erdfelder E, Lang AG, Buchner A (2007) G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods* 39:175–191

11. Faul F, Erdfelder E, Bucher A, Lang AG (2009) Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav Res Methods* 41:1149–1160
12. R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>. Accessed September 2016
13. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575
14. Gogarten SM, Bhangale T, Conomos MP, Laurie CA, McHugh CP, Painter I, Zheng X, Crosslin DR, Levine D, Lumley T, Nelson SC, Rice K, Shen J, Swarnkar R, Weir BS, Laurie CC (2012) GWASTools: an R/bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* 28:3329–3331
15. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. *Nat Genet* 39:906–913
16. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 23:1294–1296
17. Aulchenko YS, Karssen LC (2015) The GenABEL project developers. The GenABEL Tutorial Zenodo; doi:<https://doi.org/10.5281/zenodo.19738>
18. Nicolazzi EL, Marras G, Stella A (2016) SNPConvert: SNP array standardization and integration in livestock species. *Microarrays* 5:17
19. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909
20. Rice TK, Schork NJ, Rao DC (2008) Methods for handling multiple testing. *Adv Genet* 60:293–308
21. Panagiotou OA, Ioannidis JPA, the Genome-Wide Significance Project (2012) What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int J Epidemiol* 41:273–286
22. De S, Pedersen BS, Kechris K (2014) The dilemma of choosing the ideal permutation strategy while estimating statistical significance of genome-wide enrichment. *Brief Bioinform* 15:919–928
23. Backes C, Rühle F, Stoll M, Haas J, Frese K, Franke A, Lieb W, Wichmann HE, Weis T, Kloos W, Lenhof HP, Meese E, Katus H, Meder B, Keller A (2014) Systematic permutation testing in GWAS pathway analyses: identification of genetic networks in dilated cardiomyopathy and ulcerative colitis. *BMC Genomics* 15:622
24. Turner SD (2014) Qqman: an R package for visualizing GWAS results using Q–Q and manhattan plots. *bioRxiv*. <https://doi.org/10.1101/005165>
25. Duggal P, Gillanders EM, Holmes TN, Bailey-Wilson JE (2008) Establishing an adjusted p-value threshold to control the family-wide type I error in genome wide association studies. *BMC Genomics* 9:516
26. Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* 76:449–462
27. Zheng-Bradley X, Flicek P (2016) Applications of the 1000 genomes project resources. *Brief Funct Genomics* 16(3):163–170. [Epub ahead of print] PMID: 27436001
28. Browning BL, Browning SR (2016) Genotype imputation with millions of reference samples. *Am J Hum Genet* 98:116–126
29. Aulchenko YS, Struchalin MV, van Duijn CM (2010) ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* 11:i34
30. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from next-generation sequencing data. *Nucleic Acids Res* 38:e164
31. Chang X, Wang K (2012) wANNOVAR: annotating genetic variants for personal genomes via the web. *J Med Genet* 49:433–436
32. Yang H, Wang K (2015) Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc* 10:1556–1566
33. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F (2016) The ensembl variant effect predictor. *Genome Biol* 17:122
34. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila Melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6:80–92
35. Smedley D, Jacobsen JO, Jäger M, Köhler S, Holtgrewe M, Schubach M, Siragusa E,

- Zemojtel T, Buske OJ, Washington NL, Bone WP, Haendel MA, Robinson PN (2015) Next-generation diagnostics and disease-gene discovery with the exomiser. *Nat Protoc* 10:2004–2015
36. Cheng YC, Hsiao FC, Yeh EC, Lin WJ, Tang CY, Tseng HC, Wu HT, Liu CK, Chen CC, Chen YT, Yao A (2012) VarioWatch: providing large-scale and comprehensive annotations on human genomic variants in the next generation sequencing era. *Nucleic Acids Res* 40(Web Server issue):W76–W81
 37. Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, Lee BT, Learned K, Karolchik D, Hinrichs AS, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Fujita PA, Eisenhart C, Diekhans M, Clawson H, Casper J, Barber GP, Haussler D, Kuhn RM, Kent WJ (2016) The UCSC genome browser database: 2016 update. *Nucleic Acids Res* 44 (D1):D717–D725
 38. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Keenan S, Lavidas I, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Nuhn M, Parker A, Patricio M, Pignatelli M, Rahtz M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Birney E, Harrow J, Muffato M, Perry E, Ruffier M, Spudich G, Trevanion SJ, Cunningham F, Aken BL, Zerbino DR, Flicek P (2016) Ensembl 2016. *Nucleic Acids Res* 44 (D1):D710–D716
 39. Shabalin AA (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28:1353–1358
 40. Claussnitzer M, Dankel SN, Klocke B, Grallert H, Glunk V, Berulava T, Lee H, Oskolkov N, Fadista J, Ehlers K, Wahl S, Hoffmann C, Qian K, Rönn T, Riess H, Müller-Nurasyid M, Bretschneider N, Schroeder T, Skurk T, Horsthemke B, Spieler D, Klingenspor M, Seifert M, Kern MJ, Mejhert N, Dahlman I, Hansson O, Hauck SM, Blüher M, Arner P, Groop L, Illig T, Suhre K, Hsu YH, Mellgren G, Hauner H, Laumen H, DIAGRAM+Consortium (2014) Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. *Cell* 156:343–358
 41. Li MX, Gui HS, Kwan JS, Sham PC (2011) GATES: a rapid and powerful gene-based association test using extended Simes procedure. *Am J Hum Genet* 88:283–293
 42. Van der Sluis S, Dolan CV, Li J, Song Y, Sham P, Posthuma D, Li MX (2015) MGAS: a powerful tool for multivariate gene-based genome-wide association analysis. *Bioinformatics* 31:1007–1015

Whole Genome Library Construction for Next Generation Sequencing

Jonathan J. Keats, Lori Cuyugan, Jonathan Adkins, and Winnie S. Liang

Abstract

With the rapid evolution of genomics technologies over the past decade, whole genome sequencing (WGS) has become an increasingly accessible tool in biomedical research. WGS applications include analysis of genomic DNA from single individuals, multiple related family members, and tumor/normal samples from the same patient in the context of oncology. A number of different modalities are available for performing WGS; this chapter focuses on wet lab library construction procedures for complex short insert WGS libraries using the KAPA Hyper Prep Kit (Kapa Biosystems), and includes a discussion of appropriate quality control measures for sequencing on the Illumina HiSeq2000 platform. Additional modifications to the protocol for long insert WGS library construction, to assess structural alterations and copy number changes, are also described.

Key words Whole genome sequencing, Next generation sequencing, Short insert whole genome sequencing, Long insert whole genome sequencing

1 Introduction

Since the completion of the Human Genome Project [1], sequencing technology has evolved from labor-intensive and time-consuming capillary-based sequencing [2, 3] to massively parallel next generation sequencing (NGS) [4–6]. As a result of these technological advancements, WGS now enables us to spell out the entire 3.1 billion nucleotides sequence of the human genome in a dramatically more cost-effective and timely manner. Biomedical research, particularly in the area of oncology, has benefitted from such development, as WGS allows comparison between tumor genomes and the corresponding normal, or constitutional, genome to identify tumor-specific somatic alterations.

WGS encompasses a number of distinct modalities. The most commonly used approach comprises sequencing of short inserts (approximately 250–300 bp long). However, additional approaches, including mate pair sequencing, long insert WGS (inserts that are

~850–1000 bp long), and long range sequencing (LRS), may be utilized to identify structural aberrations and variants, with the latter also enabling phasing analyses. In this chapter, we focus on construction of both short and long insert [7] WGS libraries.

At present, the most widely adopted sequencing platforms are those developed by Illumina (San Diego, CA). The Illumina platform utilizes a sequencing-by-synthesis (SBS) chemistry based on proprietary reversible dye-terminator nucleotide analogues [8, 9]. Numerous companies sell whole genome library preparation kits to generate adapter-ligated libraries for sequencing on Illumina platforms. Major differences among kits include variable techniques for fragmentation and adapter ligation, DNA input amounts, adapters/indexing formats, and enzymes used for polymerase chain reaction (PCR)- enrichment of libraries, if enrichment is performed. Here we describe a foundation protocol used for generating short insert WGS libraries using KAPA Hyper Prep Kit (Kapa Biosystems), and include modifications for generating long insert WGS libraries [7].

The KAPA Hyper Prep kit was selected for this protocol based on (1) an experimental comparison of the Kapa Biosystems Library Preparation kit using XT2 adapters (Agilent) against the XT2 Library Prep kit (Agilent) and the Ultra DNA Library Prep Kit (New England Biolabs); (2) evaluation of ligation efficiency between Agilent XT and XT2 adapters when using the KAPA Hyper Prep Kit; and (3) evaluation of starting DNA to library molecule conversion efficiency using the on-bead protocol versus the Hyper Prep protocol (*see* Chapter 10). As the construction of an exome library begins with the generation of a whole genome library prior to target enrichment, analysis of pre-capture whole genome libraries was performed. Construction of short insert WGS libraries generates DNA fragments that are approximately 250 bp in size, following incorporation of 123 bp indexed adapters, whereas construction of whole genome libraries for whole exome capture typically generates 150–200 bp DNA fragments and may incorporate ligation of short non-indexed adapters prior to bait hybridization. In comparative analyses utilizing short non-barcoded adapters for whole genome library construction during exome library preparation, we observed dramatic differences in performance among the three kits (*see* Chapter 10).

2 Materials

2.1 Library Preparation Reagents, Consumables, and Additional Items

1. KAPA Hyper Prep Kit (Kapa Biosystems).
2. Non-indexed XT adapters (Agilent); XT2 adapters (Agilent) for long insert WG libraries.
3. Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific).
4. Pipette tips.

5. Reagent reservoirs.
6. Lo-Bind 1.5 mL Eppendorf Tubes (VWR).
7. 95 microTUBE plate (for E220) or Snap Cap microTUBE (for S220) (Covaris).
8. 96-well 0.3 mL PCR plates.
9. PCR plate seals.
10. Molecular grade absolute ethanol.
11. Molecular grade water.
12. 50 mL conical tubes for preparing ethanol washes.
13. 1× TElowE buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH = 8.0).
14. Agencourt AMPure XP Beads (Beckman Coulter).
15. High Sensitivity D5000 ScreenTape and reagents for long insert WG libraries (Agilent).
16. High Sensitivity D1000 ScreenTape and reagents (Agilent).
17. D1000 ScreenTape and reagents (Agilent).
18. Pippin Prep pre-cast gel cassettes (Sage Biosciences).
19. Ice bucket and ice.

2.2 Accessories/ Instruments

1. Pipettes (single and multichannel).
2. Qubit 3.0 Fluorometer (Thermo Fisher Scientific).
3. Ring magnet plate (Beckman Coulter Genomics).
4. Benchtop vortexer.
5. Thermal cycler with a 0.2 mL heat block and heated lid.
6. E220 or S220 Focused-ultrasonicator (Covaris).
7. 2200 TapeStation Instrument (the Bioanalyzer instrument [Agilent] may also be used).
8. Pippin Prep (Sage Biosciences).

3 Methods

For short insert WGS library construction, the manufacturer's protocol for the KAPA Hyper Prep Kit can be followed. As the protocol provides ranges for various parameters, suggested parameters outside of the manufacturer's protocol are described below, along with additional details not provided in the protocol. Recommended quality control (QC) measures are shown in *italics*. Unless modifications are otherwise stated (marked by asterisks), the protocol for long insert WGS library construction follows the manufacturer's protocol for the KAPA Hyper kit along with short insert WGS recommendations below. An overview of the short and long insert WGS library preparation protocols is shown in Fig. 1.

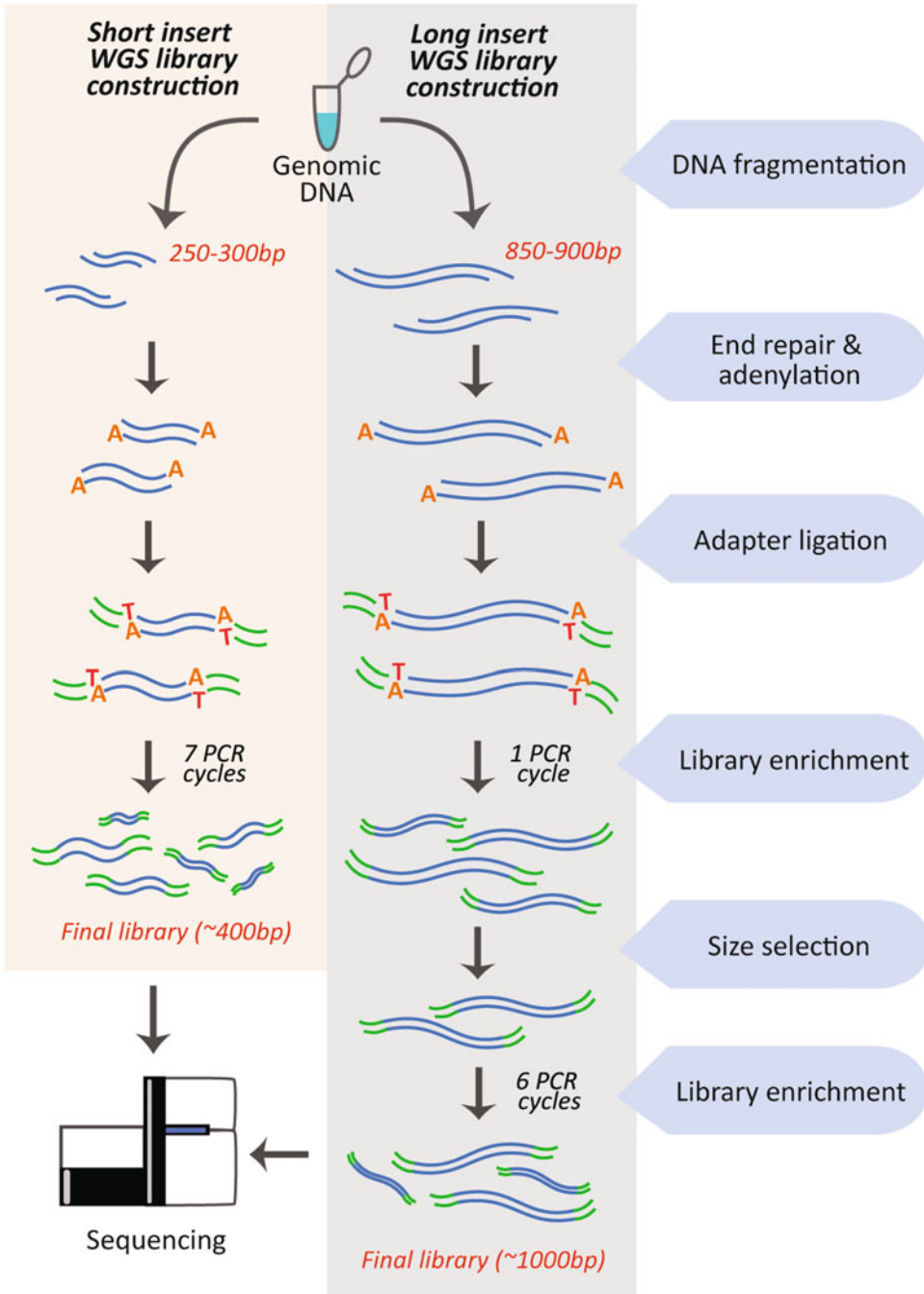


Fig. 1 Short versus long insert WGS library construction. A comparison of short versus long insert WGS library generation protocols is shown. Key protocol differences include DNA fragmentation and library enrichment, as well as an additional size selection step for constructing long insert WGS libraries

3.1 DNA Fragmentation

QC: a minimum DIN (DNA Integrity Number) of 7 is recommended for DNA samples. Special focus should be placed on the ratio of total nucleic acid quantity, as measured by a spectrophotometer, versus double-stranded DNA (Qubit or TapeStation), as these library preparations require double-stranded DNA (ratios ≥ 0.6 are recommended). Final results can also be affected by DNA impurities, which can be detected by 260/280 (1.8–1.95 optimal) and 260/230 (1.9–2.2 optimal) ratios.

1. If high quality DNA is available, an input of 200 ng of double-stranded DNA into library construction is sufficient (increased inputs are suggested for degraded DNA [i.e., $DIN < 7$]). A total of 220 ng is recommended for DNA fragmentation to account for loss of material during this step and for size verification following fragmentation.

*For long insert WG libraries, 220 ng of intact DNA can be used. Degraded DNA ($DIN < 7$) is not recommended for long insert WG library construction.

2. Dilute 220 ng of DNA of each sample with TElowE buffer to a final volume of 55 μ L. DNA should always be fragmented in an EDTA-containing buffer to prevent sonication-based artifacts during sequencing.
3. The Covaris E220 fragmentation parameters to generate approximately 250–300 bp molecules follows:
 - (a) Duty cycle: 10%
 - (b) Peak Power: 175
 - (c) Cycles/burst: 200
 - (d) Time: 40 s
 - (e) Temp max: 7 °C

*Covaris E220 parameters for long insert WG library construction:

- (a) Duty cycle: 4%
 - (b) Peak Power: 170
 - (c) Cycles/Burst: 200
 - (d) Time: 20 s
 - (e) Temp max: 7 °C
4. Following fragmentation, size verification can be performed using 5 μ L (20 ng) of sample on the TapeStation using D1000 ScreenTape, reagents, and ladder. An alternative option for size verification is to electrophoretically separate each sample on a 2% TAE (Tris–acetate–EDTA) gel (1 h, 120 V) to ensure that the majority of molecules fall within the expected size ranges.

* For size verification of long insert WGS preparations, High Sensitivity D5000 ScreenTape and reagents may be used, or alternatively, 5 μL of each sample can be electrophoresed on a 1.5% TAE gel (1 h, 120 V).

3.2 End Repair and Adenylation

1. To prepare for purification, remove AMPure beads from cold storage and equilibrate to room temperature for 30 min.

3.3 Adapter Ligation and Purification

1. 5 μL of a 30 μM adapter stock is recommended for adapter ligation of each sample to generate a 1:300 molar insert ends to adaptor ends ratio (this is significantly higher than conventional preparations suggesting 1:10 ratios).

*For long insert WGS libraries, 5 μL of 6 μM of Agilent XT2 adapter stock is recommended.

2. Increasing the ligation reaction time will result in an increase in the number of ligated molecules. Fifteen minutes is sufficient for 200 ng inputs, but a longer reaction time is recommended for low input amounts, e.g., ≤ 100 ng.
3. During purification of adapter ligated molecules, a 1:1.8 (sample-AMPure XP bead) ratio is suggested. The recommended purification protocol is as follows:
 - (a) Prepare 80% ethanol using absolute ethanol and molecular grade water.
 - (b) Add 198 μL of AMPure XP beads to each 110 μL sample and pipet up and down 10 \times to generate a well-distributed suspension. Incubate at room temperature for 15 min for DNA-bead binding.
 - (c) Place sample plate on magnet and incubate at room temperature for 5 min. The liquid will be clear after 5 min. Leave plate on magnet until indicated.
 - (d) Remove and discard supernatant using a pipette. Be sure to not disturb the beads.
 - (e) Add 200 μL of 80% ethanol (do not pipet to mix). Incubate at room temperature for 30 s.
 - (f) Remove and discard supernatant using a pipette. Add an additional 200 μL of 80% ethanol (do not pipet to mix) and incubate at room temperature for 30 s.
 - (g) Remove and discard supernatant. Incubate plate at room temperature for 5 min to allow beads to dry.
 - (h) Remove plate from magnet and place on bench top. Add 22.5 μL molecular water and pipet 10 \times to mix. Incubate at room temperature for 2 min.
 - (i) Return plate to magnet and incubate at room temperature for 5 min.
 - (j) Transfer 20 μL of the supernatant, which contains adapter-ligated molecules, to a new well.

*For long insert WGS libraries, a 1:0.8 (sample: AMPure XP bead) ratio is recommended. Instead of adding 198 μL of AMPure XP beads as described above, add 88 μL of beads. All other steps described above may be followed with the exception of eluting the final product with 22.5 μL of TElowE.

4. Completion of the purification step is a safe stopping point. At this point, ligated DNA may be stored at $-20\text{ }^{\circ}\text{C}$ for up to 7 days.

3.4 Library Enrichment and Purification

1. For the enrichment cycling protocol, use an optimized primer annealing temperature of $65\text{ }^{\circ}\text{C}$ for 15 s. Seven PCR cycles is sufficient for enrichment.

*For long insert WGS libraries, a pre-size selection enrichment step is performed using the same library amplification reagent volumes as for short insert WGS libraries and one cycle of PCR using the following program:

- (a) $98\text{ }^{\circ}\text{C}$ for 60 s
- (b) $63\text{ }^{\circ}\text{C}$ for 30 s
- (c) $72\text{ }^{\circ}\text{C}$ for 60 s
- (d) $72\text{ }^{\circ}\text{C}$ for 2 min
- (e) $4\text{ }^{\circ}\text{C}$ hold

2. During purification of enriched library molecules, add 90 μL of AMPure XP beads directly to each sample and pipet $10\times$ to mix. Incubate at room temperature for 15 min. Follow **steps 3c** through 3g in Subheading 3.3 above. Remove plate from magnet and place on bench top. Resuspend in 27.5 μL molecular grade water or TElowE, and pipet $10\times$ to mix. Incubate at room temperature for 2 min. Return plate to magnet and incubate at room temperature for 5 min. Collect 25 μL of the supernatant and place in a fresh well.

*For purification of long insert WGS libraries, a 1:0.8 ratio (sample: AMPure XP beads) can be used—this entails addition of 40 μL of beads to the 50 μL PCR reaction. Following addition of beads, the mixture is incubated at room temperature for 10 min. Additional ethanol washes are performed according to Subheading 3.4, **step 2**, with the exception that final products are resuspended in 32.5 μL TElowE. The final volume of 30 μL of supernatant is then transferred to a fresh well or tube.

QC: Assess each short insert WGS library using High Sensitivity D1000 ScreenTape, reagents, and ladder. Final sequencing-ready, short insert WG libraries will be approximately 400–450 bp long. Quantify double-stranded molecules in each library using the Qubit. Generation of at least 1000 pM of library will be sufficient for downstream sequencing. To generate long insert WGS libraries, disregard the TapeStation QC step and proceed to the remaining sections.

3.5 Long Insert WG Size Selection and Purification

(This section onward is composed of additional steps for generating long insert WGS libraries only)

1. Using the Pippin prep, follow instructions outlined for the “Pippin Prep Quick Guide 1.5% Agarose Gel Cassette.” Create a narrow range protocol with collection at a target size of 975 bp (start = 925 bp, end = 1025 bp). Use 1.5% DF Marker K and ensure both the loading solution and buffer are at room temperature.
2. On the protocol editor tab, select the protocol, and follow the Quick Guide for calibrating optics, inspecting the gel cassette, and performing the continuity test (The optics should be calibrated once a day and the continuity test will fail the first time it is run each day. Any variations in temperature of the gel will affect the test).
3. Add 11 μL of loading solution/marker K mix, equilibrated to room temperature, to each sample. Mix well by pipetting 15–20 \times .
4. Follow the loading procedures in the Quick Guide. The gel will run for approximately 45 min.
5. After the run is complete, measure the recovery volume for each sample to calculate the volume of beads to be used for purification of the sample.
6. For purification of size-selected molecules, use a 1:0.8 (sample: AMPure XP bead) ratio. For example, for a 50 μL sample, add 40 μL of beads. Follow Subheading 3.4, **step 2*** with the exception of resuspending the dried pellet in 22.5 μL TElowE. The resuspension is incubated at room temperature for 2 min, transferred to the magnet for 5 min to allow for separation of beads from the supernatant, and finally 20 μL of the supernatant is transferred to a new tube. This is a safe stopping point, and size-selected DNA may be stored at $-20\text{ }^{\circ}\text{C}$ for up to 7 days.

3.6 Long Insert WG Library Enrichment and Purification

1. Following size selection, an additional library enrichment step is performed for long insert WGS libraries. The reaction is composed of:
 - (a) 2 \times KAPA HiFi master mix—25 μL
 - (b) 10 \times Kapa PCR primers—5 μL

Make a master mix of the reaction components, including 10% overflow, and place on ice, and then add 30 μL to each 20 μL sample of adapter-ligated, size-selected library. Pipet 10 \times to mix well and briefly centrifuge samples to collect the entire volume. The following cycling protocol can be used:

- (a) 98 °C for 45 s
- (b) 6 cycles of:
 - 98 °C for 15 s
 - 63 °C for 30 s
 - 72 °C for 60 s
- (c) 72 °C for 2 min
- (d) 4 °C hold

The number of cycles may be increased or decreased, depending on the amount of input DNA used for library construction.

2. For purification of final long insert WGS libraries, a 1:1 (sample: AMPure XP bead) ratio can be used. Follow Subheading 3.4, **step 2***, with the exception of resuspending the dried pellet in 27 μ L TElowE. The resuspension is incubated at room temperature for 2 min, transferred to the magnet for 5 min to allow for separation of beads from the supernatant, and then 25 μ L of the supernatant is transferred to a fresh Lo-Bind tube.

QC: Assess each library using High Sensitivity D5000 Screentape, reagents, and ladder. Final sequencing-ready long insert WGS libraries will be approximately 1000 bp long. Quantify double-stranded molecules in each library using the Qubit. Construction of libraries with concentrations greater than 1000 μ M will be sufficient for downstream sequencing. An example TapeStation trace of a final long insert WGS library is shown in Fig. 2.

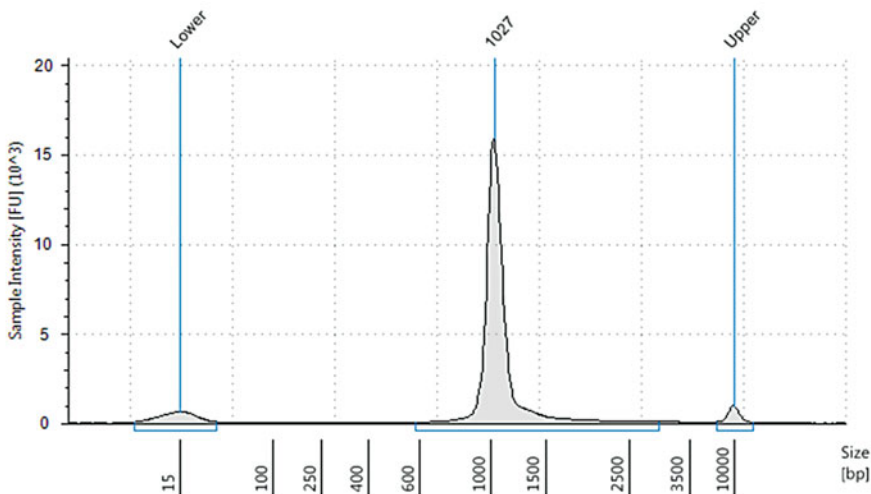


Fig. 2 Trace of a long insert WGS library. An example of an Agilent High Sensitivity D5000 TapeStation trace of a final long insert WGS library is shown. The *lower* and *upper* TapeStation markers are shown at approximately 15 and 10,000 bp, respectively, along with the completed library that is approximately 1027 bp in size. Successful library construction results in a clear defined peak

3.7 Sequencing Loading Concentrations

1. While optimization of loading concentrations of libraries should be separately performed, for short insert WGS libraries, a recommended 17 pM of library, quantified by Qubit, may be used as a target for clustering onto a lane of a V3 flowcell. For long insert WG libraries, a recommended 12 pM of library, quantified by Qubit, may be used as a target for clustering onto a lane of a V3 flowcell.

4 Notes

1. *DNA extraction:* During extraction of genomic DNA for WGS library construction, it is recommended that each sample be treated with RNase I to remove contaminating transcripts. Genomic DNA should also be accurately assessed and quantified prior to constructing libraries. The use of degraded DNA will negatively impact the outcome of library construction, and is not recommended for generation of long insert WGS libraries under any circumstances. For precious samples with significant degradation (i.e., DIN < 7), increasing input amounts may be attempted. In these cases, a decrease in the number of PCR cycles during enrichment is recommended.
2. *Protocol timeline:* Construction of short and long insert WGS libraries, along with QCs, can be completed in 1.5 and 2 days, respectively.
3. *Alternative protocols:* under circumstances in which large amounts (i.e., ≥ 1 μg) of high quality genomic DNA is available, PCR-free whole genome protocols and kits are recommended to reduce any bias that may be introduced during library enrichment. The use of UMIs (unique molecular identifiers) for non-PCR-free approaches may also be considered (*see* the Notes section in Chapter 10).
4. *Final library concentration:* if preparing libraries from degraded samples, it is likely that final library concentrations will be low. While we typically expect to construct final libraries with concentrations >1000 pM for high quality samples, sequencing may still be performed for lower yields by lowering the volume of the denaturation reaction prior to clustering.
5. *Tumor/normal whole genome analysis:* for generation of paired tumor/normal whole genomes for identification of somatic alterations, a number of considerations should be addressed:
 - (a) For sequencing, higher coverage of the tumor whole genome compared to the normal whole genome is recommended to capture potential sub-clonal tumor populations and to overcome potential contaminating normal cells.

- (b) Tumor cellularity estimates are often performed for tumor biopsies by a board-certified pathologist. While variability in these estimates is frequently observed compared to true estimates based on mutation allele frequencies in sequencing data, a priori knowledge of estimates may be used to guide determination of sequencing depth, e.g., samples with lower tumor cellularity estimates may be sequenced to higher depths.
- (c) For somatic analysis of tumor/normal WGS, reference controls may be prepared and sequenced alongside experimental samples. Using this strategy, known point or insertion/deletion (indels) mutations may be used as references for somatic variant calling. A PCR-free somatic reference generated from a matched metastatic melanoma cell line (COLO829) and normal was recently constructed across multiple institutions for this purpose [10].

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
2. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74:5463–5467
3. Swerdlow H, Wu SL, Harke H, Dovichi NJ (1990) Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *J Chromatogr* 516:61–67
4. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet* 24:133–141. <https://doi.org/10.1016/j.tig.2007.1012.1007>. Epub 2008 Feb 1011
5. Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26:1135–1145. <https://doi.org/10.1038/nbt1486>
6. Tucker T, Marra M, Friedman JM (2009) Massively parallel sequencing: the next big thing in genetic medicine. *Am J Hum Genet* 85:142–154. <https://doi.org/10.1016/j.ajhg.2009.1006.1022>
7. Liang WS, Aldrich J, Tembe W, Kurdoglu A, Cherni I, Phillips L, Reiman R, Baker A, Weiss GJ, Carpten JD et al (2014) Long insert whole genome sequencing for copy number variant and translocation detection. *Nucleic Acids Res* 42:e8. <https://doi.org/10.1093/nar/gkt1865>. Epub 2013 Sep 1025
8. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59. <https://doi.org/10.1038/nature07517>
9. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res* 34:e22
10. Craig DW, Nasser S, Corbett R, Chan SK, Murray L, Legendre C, Tembe W, Adkins J, Kim N, Wong S et al (2016) A somatic reference standard for cancer genome sequencing. *Sci Rep* 6:24607. <https://doi.org/10.1038/srep24607>

Whole Exome Library Construction for Next Generation Sequencing

**Winnie S. Liang, Kristi Stephenson, Jonathan Adkins,
Austin Christofferson, Adrienne Helland, Lori Cuyugan,
and Jonathan J. Keats**

Abstract

Whole exome sequencing (WES) is a DNA sequencing strategy that provides a survey of base substitutions across coding genomic locations and other regions of interest. As the coding portion of the genome encompasses only 1–2% of the entire genome, this approach represents a more cost-effective strategy to detect DNA alterations that may alter protein function, compared to whole genome sequencing. Although the research community has and is currently delineating the functional implications of sequence changes in noncoding regions of the genome, WES is a currently available assay that provides valuable information for both discovery research and precision medicine applications. In this chapter, we present a WES library preparation protocol using the KAPA Hyper Prep Kit with Agilent SureSelect Human All Exon V5+UTR probes that demonstrates high DNA-to-library conversion efficiency for sequencing on the Illumina HiSeq platform.

Key words Whole exome sequencing, Next generation sequencing, DNA substitutions, Rare diseases, Coding region, Library preparation

1 Introduction

Whole exome sequencing (WES) allows for the comprehensive survey of the coding regions of a genome to identify base changes, and may also be used to identify copy number alterations. As this approach entails a selection step during library construction to specifically select genomic locations using predesigned probes, additional targets may also be evaluated, including splice sites, untranslated regions (UTRs), promoters, and introns. Since one of the earliest demonstrations of exome analysis using PCR [1], WES has become widely adopted in the biomedical research communities. Pioneering efforts in both discovery research and medical genetics largely spearheaded this widespread embrace of WES technology [2–5]. In oncology, next generation sequencing has also

dramatically enabled identification of driver events and other key DNA mutations in specific cancer types and subtypes [6–11]. As a result, such events have led to the development of numerous commercially available, targeted sequencing panels, which provide a more cost-efficient approach for evaluating specific genes and mutations. Furthermore, while identification of somatic mutations in cancer can be performed through WES of both tumor DNA and the individual's constitutional DNA, germ line analysis may also be performed using WES of only an individual's constitutional DNA [12].

Due to the adoption of WES in multiple areas of research, off-the-shelf WES library preparation kits are widely available. However, the performance of these kits is highly dependent on numerous factors, including the quality of input DNA, the specific enzymes used for ligation and library enrichment, types of adapters used during ligation, etc. In this chapter, data from unbiased comparative analyses of various library preparation kits and protocols are presented. While manufacturer's protocols associated with currently available kits are often well optimized, the data presented here demonstrate advantageous features of currently available reagents, namely the KAPA Hyper Prep Kit from Kapa Biosystems, during the library generation process, when used in combination with Agilent SureSelect Human All Exon baits. We further provide additional recommendations for generating high quality WES libraries for next generation sequencing using the Illumina HiSeq platform. Results from the comparisons presented in this chapter were also utilized to outline whole genome sequencing library construction methods (described in Chapter 8).

2 Materials

2.1 Library Preparation

1. KAPA Hyper Prep Kit (Kapa Biosystems).
2. Non-indexed XT adapters (Agilent) or synthesized adapters.

As described in Subheading 3.3, adapters may be separately synthesized and purchased, e.g., by Integrated DNA Technologies, using the same sequences as the Agilent adapters. It is recommended that synthesized adapters be HPLC (high performance liquid chromatography)-purified and ordered at a concentration of 30.3 μM —5 μL of these adapters may be used for ligation to allow for a molar ratio of approximately 100:1 for exome library preparation of 200 ng of input DNA.

3. SureSelect Human All Exon V5+UTR kit (Agilent).
4. SureSelectXT Reagent Kit (Agilent).
5. Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific).
6. Pipette tips.

7. Reagent reservoirs.
8. Lo-Bind 1.5 mL Eppendorf Tubes.
9. 95 microTUBE plate (for E220) or Snap Cap microTUBE (for S220) (Covaris).
10. 96-well 0.3 mL PCR plates.
11. PCR plate seals.
12. Molecular grade absolute ethanol.
13. Molecular grade water.
14. 50 mL conical tubes for preparing ethanol washes.
15. 1× TElowE buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH = 8.0).
16. Agencourt AMPure XP Beads (Beckman Coulter).
17. High Sensitivity D1000 ScreenTape (Agilent) and reagents.
18. D1000 ScreenTape (Agilent) and reagents.
19. Ice bucket and ice.

2.2 Accessories/ Instruments

1. Pipettes (single and multichannel).
2. Qubit 3.0 Fluorometer (Thermo Fisher Scientific).
3. Ring magnet plate (Beckman Coulter Genomics).
4. Benchtop vortexer.
5. Thermal cycler with a 0.2 mL heat block and heated lid.
6. E220 or S220 Focused-ultrasonicator (Covaris).
7. 2200 TapeStation Instrument (Agilent; the Bioanalyzer may be used as an alternative).

3 Methods

3.1 Evaluation of Multiple Exome Library Preparation Methods Using Indexed Adapters

During exome library generation, different types of adapters may be ligated to end-repaired, adenylated molecules. For example, Agilent XT2 adapters are full-length *indexed* adapters that are ligated to molecules prior to library enrichment. Following ligation and enrichment, samples are then pooled before hybridization-capture of the exome. In contrast, Agilent XT adapters are short *non-indexed* adapters that are ligated to molecules prior to enrichment and exome capture. These adapters do not contain index sequences; instead, these are incorporated during library enrichment following exome capture. These kinds of adapters result in individual exome libraries that may be pooled after generation of libraries, whereas the use of the first type of adapter results in a final library containing a pool of multiple indexed exomes. In other words, indexed adapters allow for precapture pooling of libraries, whereas non-indexed adapters are used to generate single-plex

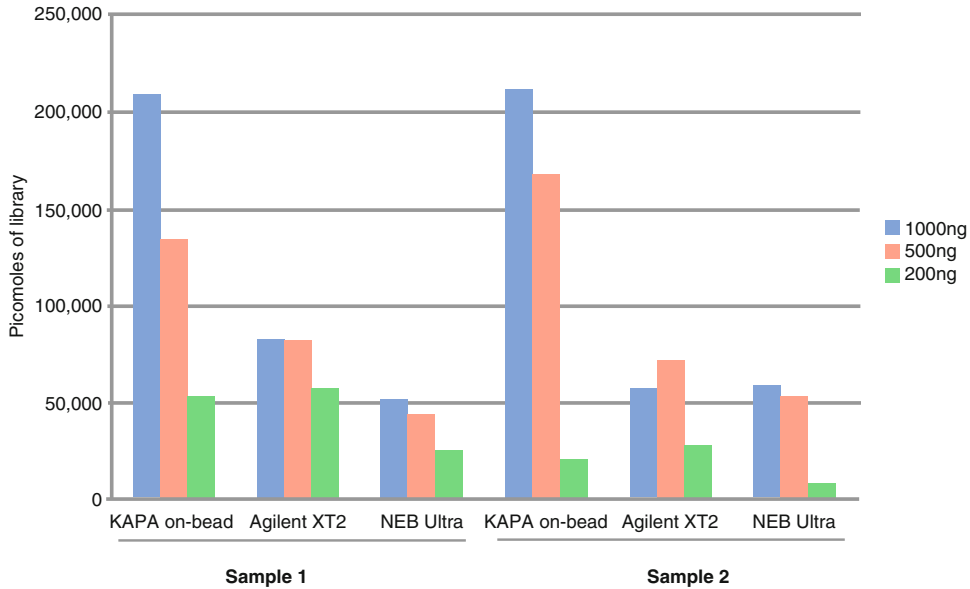


Fig. 1 Final library yields across evaluated methods. Two samples using three different input amounts for library preparation were compared across three library preparation kits. The amount of library yielded from each sample, input amount, and kit are shown

exome captures such that pools may be constructed following library construction.

For the first analysis, we compared library preparation methods utilizing the same indexed Agilent XT2 adapters and Agilent SureSelect V5+UTR capture baits. We tested the XT2 Library Preparation kit (Agilent), the KAPA Library Preparation kit using the with-bead or on-bead protocol (Kapa Biosystems), and the Ultra DNA kit (New England Biolabs). Three different starting DNA input amounts, 200 ng, 500 ng, and 1000 ng, were evaluated across all three kits. The manufacturer's protocols were followed for library construction using two separate DNA samples. Final libraries were sequenced by synthesis for 2×100 read lengths on the Illumina HiSeq2000.

We observed dramatic differences in final library yield across the three kits (Fig. 1). The KAPA Library Preparation kit using the on-bead protocol showed the greatest library yield for each input amount, with 500 ng and 1000 ng inputs resulting in the highest yields. Following sequencing, the total number of library molecules was estimated using Picard's (<http://broadinstitute.github.io/picard>) MarkDuplicates tool for each kit and input amount (Fig. 2). We found that all the KAPA libraries, across all input amounts, demonstrated the greatest number of library molecules. Notably, the 200 ng input KAPA library showing improved performance over the 500 ng and 1000 ng input libraries generated using the Agilent kit. Using the same Picard tool, the percentage of

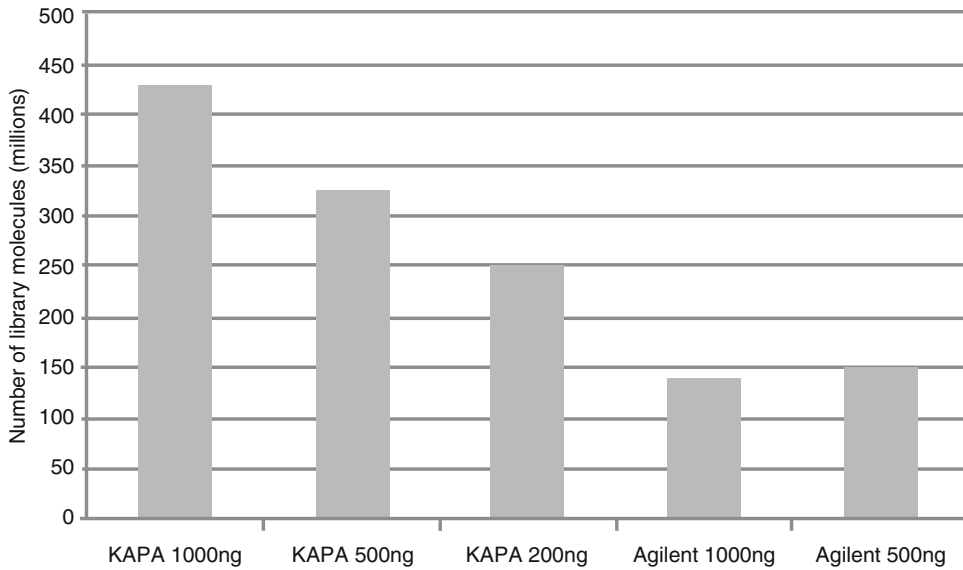


Fig. 2 Estimated number of unique library molecules. For the KAPA on-bead and Agilent XT2 library preparations, the total number of library molecules was estimated using the MarkDuplicates tool (Picard) for each input amount

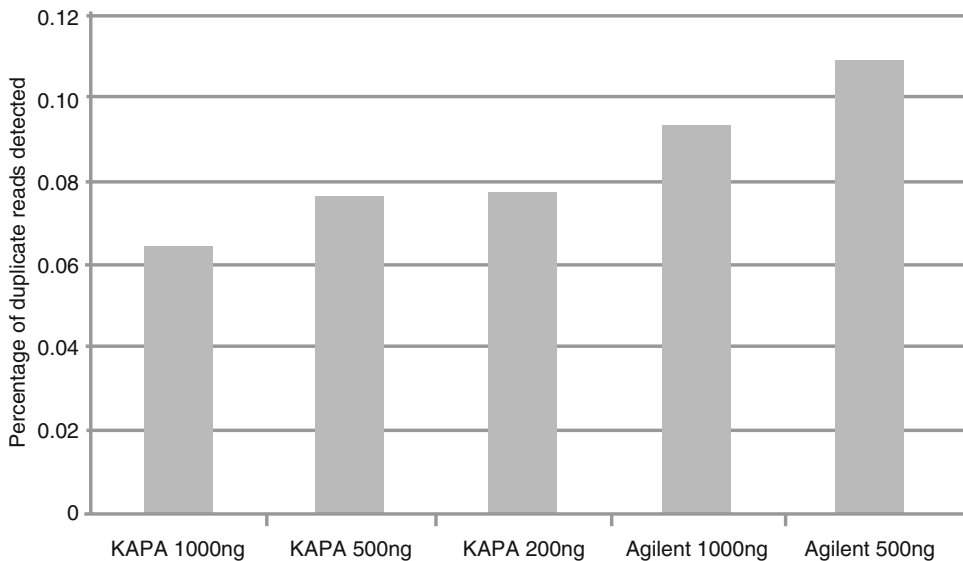


Fig. 3 Percentage of duplicate reads detected. Using the MarkDuplicates tool, the percentage of duplicated fragments was approximated using the KAPA and Agilent kits and for three different input amounts

duplicated fragments was also estimated across these same kits and input amounts (Fig. 3). Here, the 500 ng and 1000 ng input Agilent libraries demonstrated the highest level of duplication compared to all KAPA libraries. In a final analysis, the approximate under-representation of GC-rich regions was evaluated using the

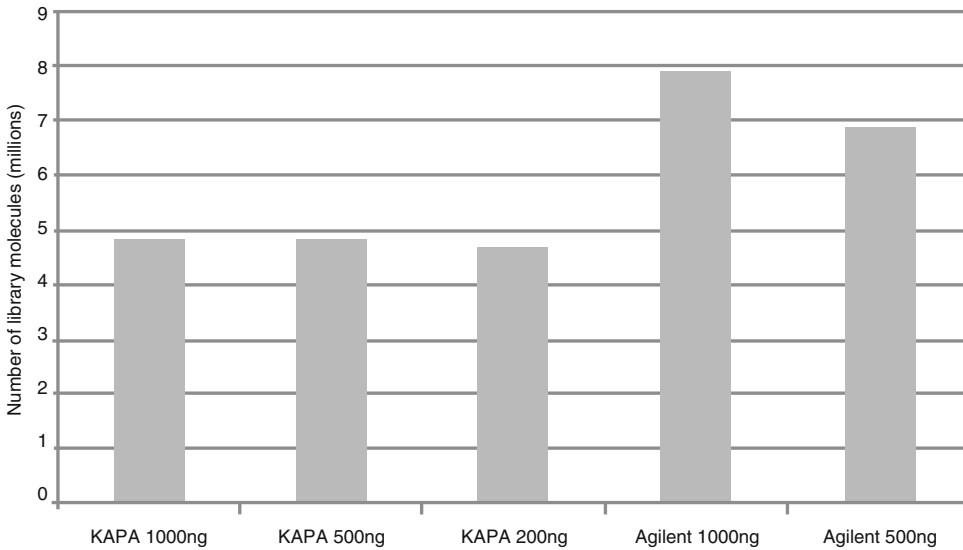


Fig. 4 Relative underrepresentation of GC-rich regions. The HSmetrics tool (Picard) was used to estimate under-representation of GC-rich regions across library preparations

Picard HSmetrics tool (Fig. 4). Both the 500 ng and 1000 ng Agilent libraries showed the greatest under-representation of GC-rich regions compared with all three KAPA libraries. It is worth noting that for the Agilent and NEB libraries, the same amount of XT2 adapter was used regardless of DNA input amount in according with the manufacturer's recommendation (i.e., 5 μ L of adapter), whereas the adapter amounts used for the KAPA libraries were adjusted linearly based on the DNA input amount (i.e., 1 μ L for 200 ng input, 2.5 μ L for 500 ng input, and 5 μ L for 1000 ng input). Overall, the KAPA Library Preparation Kit, using the on-bead protocol, shows better performance compared to the Agilent and NEB kits. Importantly, 200 ng input KAPA libraries generated about the same number of unique library molecules as Agilent libraries using 500 ng and 1000 ng input DNA.

3.2 Evaluation of Amplification Efficiencies of Whole Genome Libraries Constructed Using Indexed Adapters

Library construction is strongly influenced by amplification efficiency during library enrichment. To evaluate changes in efficiencies associated with variable DNA: adapter ratios used to determine the amount of adapter to add during ligation, as well as the number of PCR cycles used during library amplification, whole genome libraries were constructed using 200 ng input DNA, the KAPA Hyper Prep Kit, four separate DNA: Agilent XT2 adapter ratios, and either four or six PCR cycles of library enrichment (Fig. 5). Assuming that amplification efficiency remains consistent during PCR cycles, we determined the approximate amplification efficiency for the KAPA Hyper Prep Kit using indexed Agilent XT2 adapters and 200 ng of starting DNA to be 65%. While efficiency may vary across XT2 and XT adapters, we chose to describe the use

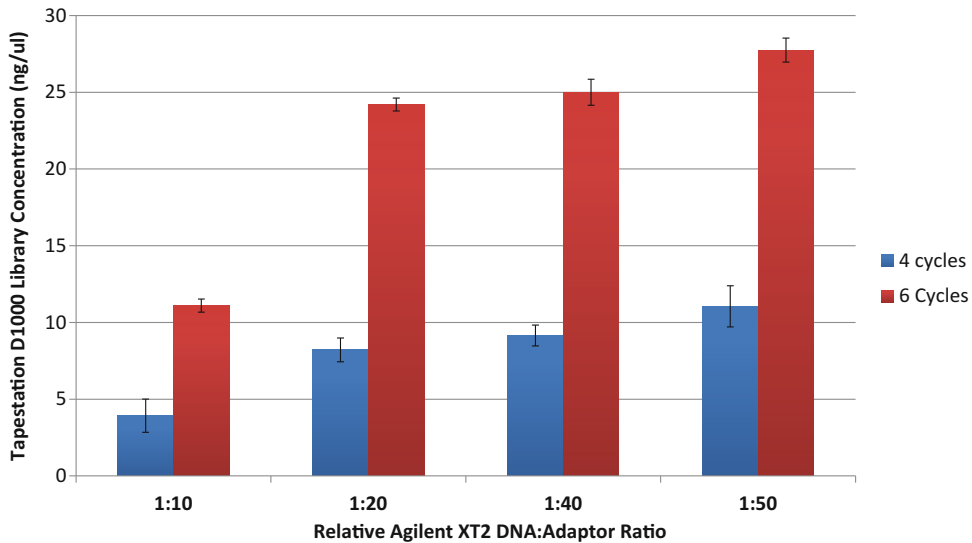


Fig. 5 Amplification efficiency of whole genome library construction using indexed adapters. To evaluate changes in efficiency associated with different DNA: adapter ratios and PCR cycles used in library enrichment, whole genome libraries were generated using 200 ng input DNA, the KAPA Hyper Prep Kit, four different DNA: adapter ratios with Agilent XT2 adapters, and either four or six PCR cycles during library enrichment

of XT adapters for the protocol shown here to allow for construction of individual single-plex exome libraries.

3.3 Evaluation of DNA-to-Library Conversion Efficiencies Across Kapa Biosystems Kits

To determine if conversion efficiencies vary depending on the Kapa Biosystems' kit used, multiple whole genome library preparations for exome construction were performed by automation on the Agilent Bravo robot using various DNA: Agilent XT adapter ratios and different enrichment primer concentrations. Library conversion efficiencies were calculated across libraries constructed using either the KAPA Library Preparation Kit and the on-bead protocol or the KAPA Hyper Prep Kit (Fig. 6). A dramatic increase in library conversion efficiency was observed when (1) the adaptor ratio and amplification primer concentrations were increased twofold and (2) the KAPA Hyper Prep Kit was used with an optimized 1:300 DNA: adapter ratio and a primer concentration twice that recommended in the XT protocol (Agilent). At this time, we now synthesize our own short adapters with a known concentration (*see* Subheading 2), instead of the adapters provided in the Agilent kits, which are used at an explicit 1:100 molar ratio of insert ends to adaptor ends.

3.4 Recommended Exome Library Construction Protocol Using the KAPA Hyper Prep Kit

Based on the results described in Subheadings 3.1–3.3, we recommend the KAPA Hyper Prep Kit with short non-indexed Agilent XT adapters for single-plex exome library preparation, using the manufacturer's protocol (*see* Notes 1–2). We chose to describe a single-plex exome library construction protocol here due to its versatility over pooled exome construction, as repooling, and

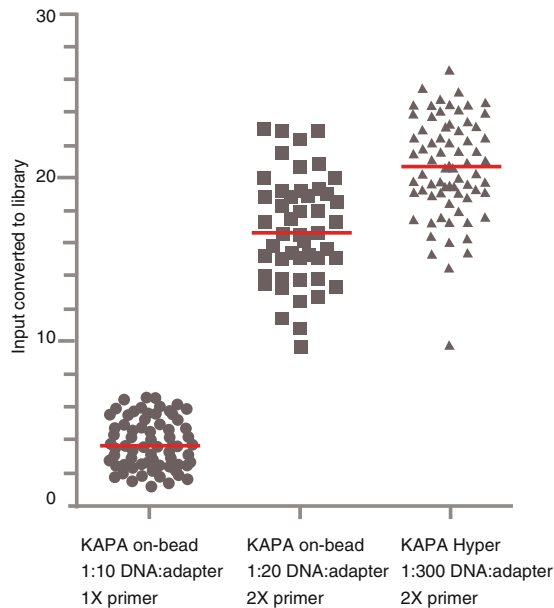


Fig. 6 Increased conversion efficiency demonstrated by the KAPA Hyper Prep Kit. Multiple whole genome library preparations for exome construction were performed using various DNA: Agilent XT adapter ratios and different enrichment primer concentrations. Library conversion efficiencies were calculated for libraries constructed using either the KAPA Library Preparation Kit (on-bead), or the KAPA Hyper Prep Kit

resequencing of libraries to address bias in index breakdowns during sequencing may be easily performed with single-plex construction. Furthermore, precapture pooled exome libraries may not yield similar amounts of individual libraries if the DNA used for library construction have variable levels of quality. For example, degraded DNA appears to be represented at a lower percentage in pools compared to higher quality DNA (*see Note 1*). Lastly, we describe hybridization capture using Agilent SureSelect baits following the SureSelect XT Target Enrichment System protocol for Illumina Multiplexed Sequencing version 1.5 (November). We also suggest the following modifications to the manufacturers' protocols:

1. For high quality DNA ($DIN > 7$), we recommend starting with 220 ng of genomic DNA to account for loss of material during both fragmentation and size verification (*see Note 3*).
2. For Covaris fragmentation optimized on an E220, the following parameters are suggested:
 - Duty cycle: 10%
 - Peak Power: 175
 - Cycles/Burst: 200
 - Time: 300 s (150 bp)
 - Temp max: 7

Fragmented DNA may be stored at 4 °C overnight or –20 °C for longer periods. Size verification can be performed using 2 µL (~8 ng) of sample on the TapeStation using the High Sensitivity D1000 ScreenTape, reagents, and ladder. An alternative option for size verification is to electrophoretically separate 5 µL (20 ng) of each sample on a 2% TAE (Tris–acetate–EDTA) gel (1 h, 120 V) to ensure that the majority of molecules fall within the expected size range (150–250 bp).

3. For adapter ligation to generate single-plex exome libraries, Agilent XT adapters should be used. Doubling the stock concentration of adapters is recommended (15–30 µM). A 1:100 DNA: adapter ratio is suggested for each reaction *see* **Note 4**.
4. Increase the annealing temperature of the XT primers to 65 °C and the number of PCR cycles to seven during the library amplification step, prior to hybridization of exome baits.
5. For AMPure XP bead purifications, the KAPA Hyper prep protocol can be followed with the exception of using a 1.8× sample volume: bead volume ratio, rather than the 0.8× ratio recommended by the KAPA protocol. This change accommodates the purification of shorter molecules (approximately 150–250 bp) generated for exome library construction, compared to whole genome library construction.
6. During post-capture enrichment, use PCR primers provided in the Agilent SureSelectXT Reagent kit.
7. For final library quantitation, the Qubit dsDNA assay may be used to quantify the amount of double-stranded library molecules present. To assess the size and distribution of each final library, they may be analyzed using D1000 ScreenTape and reagents on the Agilent TapeStation (or using Agilent Bioanalyzer reagents on the Bioanalyzer instrument). An additional option for accurate quantification of libraries is quantitative PCR, for which off-the-shelf kits are available from numerous commercial entities in the NGS space (Illumina, Agilent, Kapa Biosystems, etc.) *see* **Note 5**.

3.5 Sequencing Loading Concentrations

While optimization of loading concentrations of libraries should be performed separately, for whole exome libraries constructed following the described recommendations, 13 pM of library, quantified by Qubit, is suggested as a target for clustering onto a lane of the V3 flowcell *see* **Note 6**.

4 Notes

1. *DNA extraction*: During extraction of genomic DNA for exome library construction, it is recommended that each sample be treated with RNase I to remove contaminating RNA

transcripts. Genomic DNA should also be accurately assessed and quantified prior to beginning library construction. The use of degraded DNAs will interfere with the successful construction of sequencing libraries. For precious samples showing degradation ($DIN < 7$), or for DNA extracted from FFPE (formalin-fixed paraffin embedded) tissue, we recommend increasing DNA input amounts. Degraded FFPE DNA may be partially repaired using NEBNext FFPE DNA Repair Mix (New England Biolabs), which fills in gaps and generates blunt-ended molecules prior to fragmentation.

2. *Experimental timeline*: Construction of exome libraries, along with QC analysis, can be completed in 3–4 days.
3. *DNA fragmentation*: DNA fragmentation is the first step in all library preparation methods. The most common approaches are mechanical or enzymatic. While a number of studies have reported minor biases across various approaches [13, 14], the Covaris protocol, which uses an acoustic transducer to randomly shear DNA, has demonstrated consistent performance across libraries.
4. *Unique molecular identifiers (UMIs)*: During post-sequencing analysis, the percentage of duplicate molecules (Picard) can be used to evaluate the quality of a sequenced library. In general, a lower percentage indicates that the library is less saturated by PCR duplicates, and suggests that a sample will be more accurately represented by its sequencing data. However, it is possible that two identical fragments may represent two true library molecules that are not duplicates of one another. To address this possibility, UMI sequences [15] can be incorporated directly into adapters that are ligated to adenylated molecules. In doing so, UMIs specifically tag separate molecules, which will be separately amplified during PCR enrichment. Following alignment of sequencing data, the reads corresponding to these molecules, which may align to the same genomic location, will help to differentiate between PCR duplicates and true duplicates. As a result, less duplicate data will be discarded, subsequently increasing the amount of coverage obtained from a sample and improving the ability to accurately estimate allele ratios. With this approach, the ability to identify more rare variants and mutations is enhanced. This application is also relevant for RNA sequencing of single cells [16]. Despite concerns with sequencing errors occurring at UMIs [17], and while the use of UMIs is still being evaluated to ensure that they do not introduce additional biases, more widespread adoption is anticipated.
5. *Library yields*: if preparing libraries from degraded samples, it is expected that final library concentrations may be lower. While we typically expect to construct final libraries that are >1000

pM for high quality samples, sequencing may still be performed for lower concentrations by lowering the volume of the denaturation reaction prior to clustering.

6. *Tumor/normal whole genome analysis*: for generation of paired tumor/normal whole exomes for identification of somatic alterations in the context of cancer, a number of considerations should be addressed:
 - (a) For sequencing, higher mean coverage of the tumor exome compared to the normal exome is recommended to improve detection of low frequency changes created by sub-clonal tumor populations or contaminating normal cells in the tumor specimen.
 - (b) Tumor cellularity estimates are often performed for tumor biopsies by a board-certified pathologist prior to sequencing of a sample. While variability in these estimates is frequently observed compared to true estimates based on mutation allele frequencies in sequencing data, a priori knowledge of estimates may be used to guide determination of sequencing depth, e.g., samples with lower tumor cellularity estimates may be sequenced to higher depths.

References

1. Ley TJ, Minx PJ, Walter MJ, Ries RE, Sun H, McLellan M, DiPersio JF, Link DC, Tomasson MH, Graubert TA et al (2003) A pilot study of high-throughput, sequence-based mutational profiling of primary human acute myeloid leukemia cell genomes. *Proc Natl Acad Sci U S A* 100:14275–14280. Epub 12003 Nov 14212
2. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S et al (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* 106:19096–19101. <https://doi.org/10.11073/pnas.0910672106>. Epub 0910672009 Oct 0910672127
3. Musunuru K, Pirruccello JP, Do R, Peloso GM, Guiducci C, Sougnez C, Garimella KV, Fisher S, Abreu J, Barry AJ et al (2010) Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N Engl J Med* 363:2220–2227. <https://doi.org/10.1056/NEJMoa1002926>. Epub 1002010 Oct 1002913
4. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA et al (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 42:30–35. <https://doi.org/10.1038/ng.1499>. Epub 2009 Nov 1013
5. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE et al (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461:272–276. <https://doi.org/10.1038/nature08250>. Epub 02009 Aug 08216
6. Kanchi KL, Johnson KJ, Lu C, McLellan MD, Leiserson MD, Wendl MC, Zhang Q, Koboldt DC, Xie M, Kandoth C et al (2014) Integrated analysis of germline and somatic variants in ovarian cancer. *Nat Commun* 5:3156. <https://doi.org/10.1038/ncomms4156>
7. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, Nickerson E, Auclair D, Li L, Place C et al (2012) A landscape of driver mutations in melanoma. *Cell* 150:251–263. <https://doi.org/10.1016/j.cell.2012.1006.1024>
8. Krauthammer M, Kong Y, Ha BH, Evans P, Bacchicocchi A, McCusker JP, Cheng E, Davis MJ, Goh G, Choi M et al (2012) Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nat Genet* 44:1006–1014. <https://doi.org/10.1038/ng.2359>. Epub 2012 Jul 1029
9. The Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature* 490:61–70. <https://doi.org/10.1038/nature11251>

- doi.org/10.1038/nature11412. Epub 12012 Sep 11423
10. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, White TA, Stojanov P, Van Allen E, Stransky N et al (2012) Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet* 44:685–689. <https://doi.org/10.1038/ng.2279>
 11. Wang L, Tsutsumi S, Kawaguchi T, Nagasaki K, Tatsuno K, Yamamoto S, Sang F, Sonoda K, Sugawara M, Saiura A et al (2012) Whole-exome sequencing of human pancreatic cancers and characterization of genomic instability caused by MLH1 haploinsufficiency and complete deficiency. *Genome Res* 22:208–219. <https://doi.org/10.1101/gr.123109.123111>. Epub 122011 Dec 123107
 12. Rabbani B, Tekin M, Mahdieh N (2014) The promise of whole-exome sequencing in medical genetics. *J Hum Genet* 59:5–15. <https://doi.org/10.1038/jhg.2013.1114>. Epub 2013 Nov 1037
 13. Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D (2011) Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PLoS One* 6:e28240. [10.1371/journal.pone.0028240](https://doi.org/10.1371/journal.pone.0028240). Epub 0022011 Nov 0028230
 14. Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov MV, Oparina NY, Polozov RV, Nechipurenko YD, Grokhovskiy SL (2014) Non-random DNA fragmentation in next-generation sequencing. *Sci Rep* 4:4532. <https://doi.org/10.1038/srep04532>
 15. Kivioja T, Vaharautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, Taipale J (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* 9:72–74. <https://doi.org/10.1038/nmeth.1778>
 16. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lonnerberg P, Linnarsson S (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* 11:163–166. <https://doi.org/10.1038/nmeth.2772>. Epub 2013 Dec 1022
 17. Kou R, Lam H, Duan H, Ye L, Jongkam N, Chen W, Zhang S, Li S (2016) Benefits and challenges with applying unique molecular identifiers in next generation sequencing to detect low frequency mutations. *PLoS One* 11:e0146638. <https://doi.org/10.1371/journal.pone.0146638>. eCollection 0142016

Optimized Methodology for the Generation of RNA-Sequencing Libraries from Low-Input Starting Material: Enabling Analysis of Specialized Cell Types and Clinical Samples

Kendra Walton and Brian P. O'Connor

Abstract

RNA sequencing (RNA-seq) has become an important tool for examining the role of the transcriptome to biological processes. While RNA-seq has been widely adopted as a popular approach in many experimental designs, from gene discovery to mechanistic validation of targets, technical issues have largely limited the use of this technique to abundantly available sample sources. However, RNA-seq is becoming increasingly utilized for more specialized applications, such as flow cytometry-sorted cells and clinical specimens, due to protocol advances enabling the use of very low input material ranging from 10 pg to 10 ng of total RNA or 1–1000 intact cells. In this chapter, we present an optimized and detailed approach to RNA-seq for use with low abundance samples.

Key words RNA-seq, Ultralow abundance, mRNA

1 Introduction

Next-generation sequencing (NGS) of RNA libraries (RNA-seq) has become a powerful tool to examine transcriptional regulation in multiple biological contexts, including cell identity and lineage, cell function, tissue activity, epigenetic regulation, and disease pathogenesis [1–3]. The types of questions addressed by RNA-seq are determined at a fundamental level by the type of RNA library that is produced prior to sequencing as multiple library types exist. Here we describe the four most common types of RNA-seq applications and present a detailed protocol to perform ultralow RNA-seq (10 pg to 10 ng of total RNA or 1–1000 intact cells).

When a comprehensive study of the total RNA in a biological sample is needed, it can be accomplished via library generation using a ribosomal RNA (rRNA)-depletion approach. With this approach, rRNA is targeted for depletion while mRNA, small

noncoding RNA (sncRNA), long noncoding RNA (lncRNA), and other RNA species (i.e., bacterial or viral) in a biological sample are retained for sequencing. In general, this approach requires more starting material and a greater depth of sequencing per sample than other RNA-seq approaches, but retains the most information.

Alternatively, if mRNA is the primary target of a study, poly(A) + RNA selection can be utilized for library construction. With this approach, all poly(A)+ RNA is selected, while rRNA, sncRNA, lncRNA, and other RNA species are removed. This approach requires less material and less sequencing depth per sample compared with rRNA-seq, but at the expense of lost information with respect to the other RNA species.

In recent years, methods for library construction utilizing selection techniques for a limited number of transcripts of interest (e.g., hundreds to thousands of targets) have enabled the study of specific biological or disease pathways. These approaches require nominal amounts of input material and relatively low sequencing depth per sample, which together lead to a reduced cost per sample. This flexibility has engendered the use of targeted RNA-seq to create disease panels, which can be used in the clinical research space where sample availability is limited compared to research studies utilizing abundant sample sources, such as cell lines. However, the need for comprehensive, unbiased approaches for examining transcriptional regulation of biological and disease processes with low-input, complex specimens remains a growing need.

Researchers use RNA-seq to study complex cell types that require flow cytometry sorting for isolation, often resulting in very low total starting concentrations of RNA. This kind of application thus limits the use of traditional RNA-seq methods, such as rRNA-depletion or poly(A)+-selection [4]. Clinical research studies are also incorporating systems biology approaches to elucidate novel disease pathways and mechanisms [5]. The development of novel techniques to assess low-abundance samples, and even single cells, with RNA-seq, has transformed our ability to ask nuanced questions spanning the spectrum of basic science to the clinical research setting [6, 7]. Here we present an optimized protocol for performing RNA-seq with specialized flow cytometry-sorted cells or precious, low-abundance clinical research samples.

2 Materials

This is a two-step library build protocol that allows sequencing from picogram amounts of total input RNA. We use the SMART-Seq[®] v4 Ultra[®] Low Input RNA Kit for Sequencing and Nextera[®] XT DNA Library Prep Kit to build our libraries. The SMART-seq[®] kit generates high quality full-length cDNA from 1 to 1000 cells or 10 pg to 10 ng of total RNA in 1–10 μ l of volume.

The SMART-Seq[®] v4 Ultra[®] Low Input RNA Kit for Sequencing incorporates the use of locked nucleic acid (LNA) technology into an optimized template switching oligonucleotide. In addition, this kit can produce single-cell mRNA-seq libraries that outperform established protocols and existing kits by identifying the greatest number of transcripts. The SMART-Seq[®] v4 Ultra[®] Low Input RNA Kit for Sequencing has higher sensitivity and reproducibility than earlier versions, leading to the identification of more genes and with significantly lower background. More information on SMART technology is available on the Clontech website (<http://www.clontech.com>) [8]. The Nextera[®] XT DNA Library Prep kit fragments and tags DNA simultaneously with sequencing adapters in a single tube enzymatic reaction. Nextera[®] XT supports ultra-low DNA input of up to 1 ng to enable a wide array of input samples [9].

1. The NanoDrop spectrophotometer is used to assess RNA purity and determine which Qubit assay to use in the following step (*see Note 1*).
 - (a) NanoDrop spectrophotometer (any model).
 - (b) Pipette: 2 μ l.
 - (c) Filtered tips.
2. The Qubit assay uses a fluorescence-based dye that binds to DNA or RNA depending on the nucleic acid being measured.
 - (a) Qubit machine (any model).
 - (b) RNA Broad Range assay kit for RNA quantification.
 - (c) RNA High Sensitivity assay kit for RNA quantification.
 - (d) DNA High Sensitivity kit for DNA quantification.
 - (e) Qubit tubes.
 - (f) Pipettes: 2 μ l, 10 μ l, 200 μ l, and 1000 μ l.
 - (g) 1.7 ml eppendorf tubes or 15 ml conical tubes for master mix (depending on the number of samples to be processed).
 - (h) Filtered tips.
3. The BioAnalyzer assay gives a quantitative and qualitative assessment of the nucleic acid of interest. It is used to measure RNA before starting library generation and DNA at two different steps in the protocol.
 - (a) Agilent 2100 BioAnalyzer.
 - (b) BioAnalyzer Priming station (comes with BioAnalyzer bundle).
 - (c) Agilent RNA 6000 Nano Reagent Kit.
 - (d) Agilent RNA 6000 Pico Reagent Kit.

- (e) Agilent High Sensitivity DNA Kit.
 - (f) Pipettes: 2 μ l, 20 μ l, and 200 μ l.
 - (g) Filtered pipette tips.
 - (h) Minicentrifuge.
 - (i) IKA vortexer (comes with BioAnalyzer bundle).
4. SMART-Seq[®] v4 Ultra[®] Low Input RNA Kit for Sequencing and Nextera[®] XT DNA Library Prep Kit.
- (a) SMART-Seq[®] v4 Ultra[®] Low Input RNA Kit for Sequencing (Clontech).
 - (b) Nextera[®] XT DNA library Prep kit from Illumina (*see Note 2*).
 - (c) Single channel pipettes: 10 μ l, 20 μ l, 200 μ l, and 1000 μ l.
 - (d) Eight or twelve-channel pipettes: 10 μ l, 20 μ l, and 200 μ l.
 - (e) Sterile reagent reservoirs.
 - (f) Filter tips.
 - (g) Vortex.
 - (h) Microcentrifuge for PCR tubes and 1.5 ml tubes.
 - (i) Mini plate centrifuge or large centrifuge with buckets that will spin plates.
 - (j) 100% Molecular Biology Grade Ethanol.
 - (k) Nuclease-free water.
 - (l) Thermal cycler with heated lid.
 - (m) 96-well PCR plate or individual PCR tubes.
 - (n) Nuclease-free, low-adhesion 1.5 ml tubes.
 - (o) 15 ml or 50 ml conical tubes used for ethanol dilution and mixing.
 - (p) Adhesive plate seals.
 - (q) Magnetic stand for plates or PCR tubes.
 - (r) TruSeq Index Plate Fixture.

3 Methods

This RNA-seq protocol allows the use of small amounts of total RNA. We highly recommend assessing quality control (QC) at the front and back ends of this protocol. Store all reagents as indicated by the supplier to ensure efficiency. Be sure to work in a PCR Clean Work Station until cDNA amplification master mix is completed during the SMART-seq[®] protocol.

RNA quantification—We run three rounds of QC on samples before starting the library build.

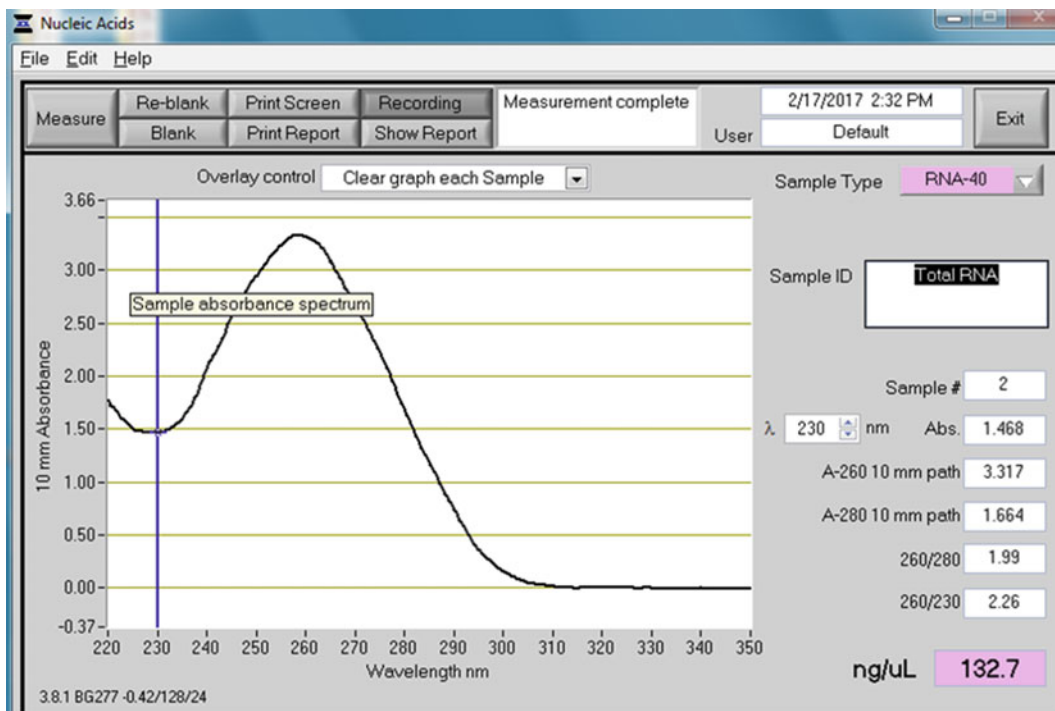


Fig. 1 Total RNA on NanoDrop. The picture shows a good total RNA trace with 1 μ l run on the NanoDrop. The 260/230 ratio is >2.0 which indicates that the RNA is “pure” and does not contain contaminants that absorb at 230 nm. The 260/280 ratio is 1.99 and indicates that the RNA is “pure” and does not contain contaminants such as phenol or proteins that absorb at 260 nm. In the presence of contaminants, both the 260/230 and 260/280 ratios would be decreased

3.1 NanoDrop Quantification

Run 1 μ l of purified total RNA on the NanoDrop spectrophotometer to estimate concentration and measure the 260/230 ratio as an estimate of nucleic acid purity. A suitable 260/230 value for pure RNA is ~ 2.0 (see Note 3, Fig. 1).

3.2 Qubit Quantification

The Qubit RNA kit is an accurate way to specifically quantitate RNA, without also measuring DNA or protein concentrations. The Qubit offers two options for measuring RNA concentration: the BR (Broad Range) and HS (High Sensitivity) kits. The Qubit BR RNA assay is used for samples between 1 ng/ μ l and 1 μ g/ μ l (see Note 4). If sample concentrations are less than 100 ng/ μ l according to NanoDrop quantitation, the Qubit RNA HS Assay kit can be used to accurately measure concentrations within the 250 pg/ μ l–100 ng/ μ l range (see Note 5). Prepare the correct Qubit RNA assay based on the estimated concentration obtained with the NanoDrop instrument. This procedure uses two known concentrations of standards against which samples are measured (see Note 6). The standards for each kit are stored at 4 $^{\circ}$ C, while all other reagents are stored at room temperature in the dark. Be sure to

keep the dye from each kit in the dark at all times, preferably by wrapping the dye tubes in foil.

1. Bring standards, which should be stored at 4 °C, to room temperature for >30 min.
2. Determine the number of samples to be measured, include the standards ($n = 2$), and add 10% to account for pipetting dead volume.

For example: 24 samples + 2 standards + 10% = 28.6 (X)

3. To make the buffer + dye master mix, use the following calculations.

$X \times 199 \mu\text{l} = \mu\text{l buffer needed}$

Example: $28.6 \times 199 \mu\text{l} = 5691.4 \mu\text{l buffer}$

$X = \mu\text{l of dye}$

Example: $X = 28.6$ (# of samples plus extra as determined above)

Add 5691.4 μl of buffer + 28.6 μl of dye, vortex to mix, and spin to collect droplets.

You can quantify 1–20 μl of sample, adjusting the volume of master mix from 199–180 μl . We use 2 μl of our sample with 198 μl of appropriate buffer–dye master mix.

4. Set out the appropriate number of tubes accounting for each sample and the two standards. Write the sample ID only on the top of the tube (*see Note 7*).
5. After mixing the Qubit master mix (buffer + dye), aliquot the appropriate amount of master mix into each sample tube (e.g., 198 μl). Aliquot 190 μl of master mix into each standard tube.
6. Add 2 μl of sample to the appropriate sample tubes. Add 10 μl of standard to the appropriate standard tube.
7. Vortex each tube for 5 s and spin to collect droplets from sides.
8. Let samples sit for 2 min before proceeding with the Qubit assay (*see Note 8*).
9. Measure the samples according to the Qubit protocol (*see Note 9*).

3.3 BioAnalyzer

The BioAnalyzer is used to assess the RNA profile and quantify RNA concentration. This assay is helpful for determining whether RNA degradation is present in samples and gives an RNA Integrity Number (RIN) based on intact 18 s and 28 s ribosomal peaks.

The RNA Nano kit detects RNA in the range of 25–500 ng/ μl . While the RNA Pico kit does not directly quantify samples, it can qualitatively assess samples in the range of 250–5000 pg/ μl . This kit is not considered suitable for quantitation unless the NanoDrop and Qubit assays fail to provide a measurement of sample concentration. The Pico kit will give a quantitative value and may sometimes be the only QC measurement obtainable from samples that

are very low in concentration. This chip also gives a RIN score for even small amounts of total RNA. In the case of ultralow protocols, as described here, we often use the qualitative and quantitative values from the Pico chip for our samples.

1. Choose the correct BioAnalyzer kit based on the RNA Qubit concentration. For the RNA Pico kit, for example, dilute samples to stay within the range specified by the kit, e.g., ~ 2 ng/ μ l, to run on the Pico chip. Remove the kit from 4 °C and let the contents come to room temperature for 30 min before use. Remove ladder from -20 °C storage, thaw on ice and prepare dilutions according to protocol. Prepare the gel matrix and aliquot into single tubes (*see Note 10*).
2. Once the kit is equilibrated to room temperature for 30 min and gel matrix had been made and aliquoted, add 1 μ l of dye to 1 aliquot of gel, vortex for 10 s, and spin at $13,000 \times g$ for 10 min. Use the gel-dye mix within 1 day.
3. Adjust the syringe clip to be compatible with the RNA chips. The silver tab should be in the topmost slot of the three available positions.
4. Put a new Pico RNA chip on the chip priming station and pipette 9 μ l of gel-dye mix slowly into the well that is marked © corresponding to the third well down on the right side (Fig. 2).
5. Be sure that the chip plunger on the chip priming station is at 1 ml, lower the lid of the chip priming station until it clicks, and push the plunger down until the top of the plunger hooks under the silver tab to hold the plunger in place.
6. Let the chip sit for exactly 30 s and release the plunger by pulling up on the silver tab, allowing the plunger to pop back up and slowly return to the 1 ml position. Let the plunger retract back toward the 1 ml position for 5 s, and then slowly pull the plunger back to the 1 ml position (*see Note 11*).
7. Once the plunger has returned to the 1 ml position, unlock the chip priming lid from the front of the chip priming station, and add 9 μ l of gel-dye mix to the two remaining “G” wells.
8. Load 9 μ l of the conditioning solution to the well that is marked “CS”.
9. Load 5 μ l of the RNA marker (green top tube) to each sample well of the chip, including the ladder position, excluding the “G” wells that already contain gel-dye mix and the well containing CS (*see Note 12*).
10. Load 1 μ l ladder into the designated well.
11. Pipette 1 μ l of sample into each sample well. If there are less samples than sample spaces, add 1 μ l of RNA marker (green top tube) to each empty well.

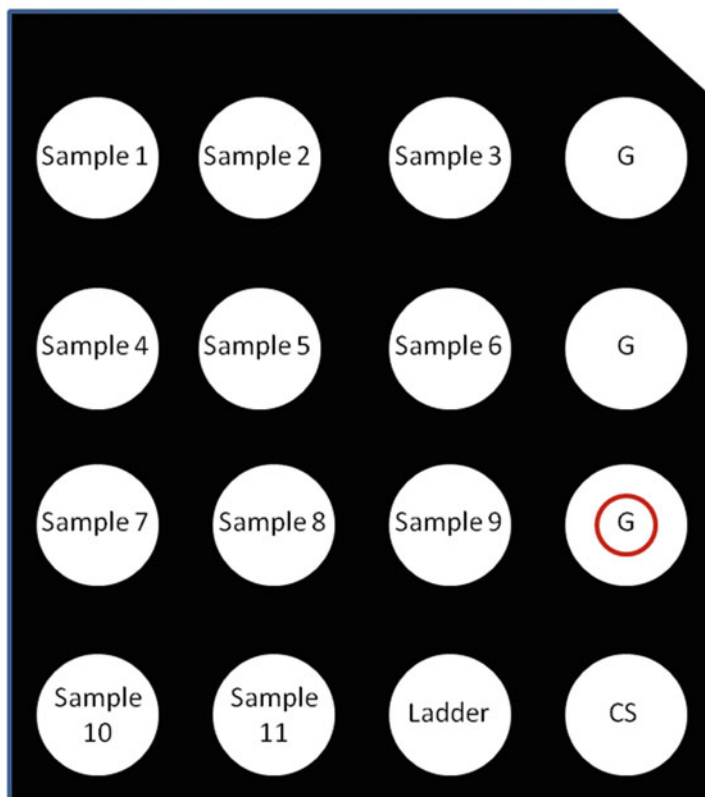


Fig. 2 Loading a Pico BioAnalyzer chip. Pipette gel–dye matrix into the 3rd well down on the right side (circled G). Once the chip is primed, add gel–dye mix to the other two wells marked “G”. Add conditioning solution to the well marked “CS”. Add marker to all sample and ladder wells. Add samples in chronological order. Add ladder to the well marked “ladder”

12. Place loaded chip horizontally onto the IKA vortexer and vortex for 1 min at 2400 rpm.
13. Run the chip on the BioAnalyzer using the Pico Eukaryote assay within 5 min.
14. Evaluate all QC results side by side. The results from the three different methods of quantification should be similar, although not necessarily identical. The RNA profile below is an example of a good RNA trace on the Pico chip (*see Note 13*, Fig. 3).

3.4 Library Preparation Using the SMART-Seq[®] v4 Ultra[®] Low Input RNA Kit for Sequencing

First Strand cDNA-Synthesis (*see Note 14*)

1. Thaw the 10× lysis buffer, RNase inhibitor, 3' SMART-Seq[®] CDS Primer IIA, SMART-Seq[®] v4 Oligonucleotide, 2× SeqAmp PCR Buffer, and PCR Primer IIA on ice. Thaw 5× Ultra Low First Strand Buffer, water, Ampure XP beads, and Elution buffer at room temperature. Leave the SMART-Scribe Reverse Transcriptase and SeqAmp DNA Polymerase at –20 °C until ready to use.

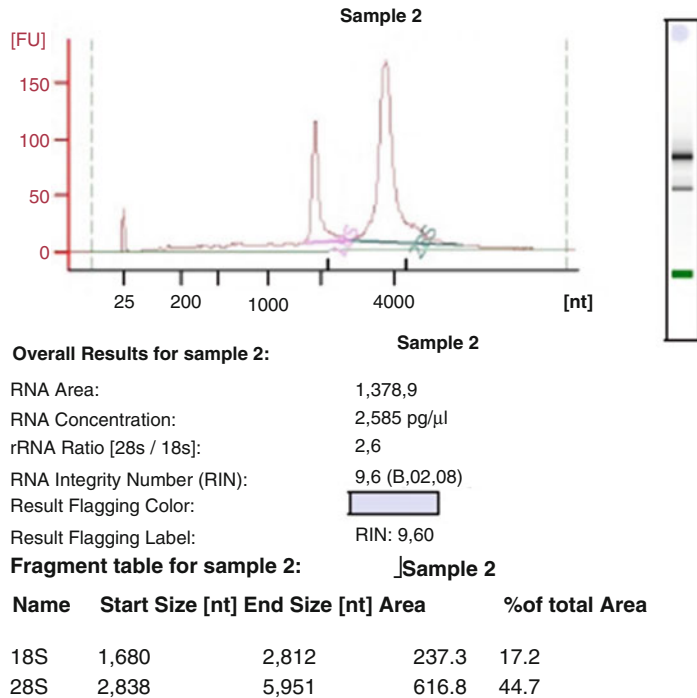


Fig. 3 Sample RNA trace. A total RNA trace on the Pico BioAnalyzer chip shows two distinct peaks at 18 s and 28 s for eukaryotes. The concentration of the sample is within range of the chip and the RIN score is 9.6, indicating intact total RNA with little to no degradation

2. Determine the amount of total RNA (or number of cells) to be used. Because concentrations are normally low for this protocol, we use the Pico BioAnalyzer values to determine starting volume and concentration of samples. For this protocol, we use 3 ng of total RNA (*see Note 15*). A positive control (*see Note 16*), which is included in the kit and a negative control should also be assayed. Bring the amount of RNA needed to 9.5 μ l in molecular biology grade water in plate or PCR tube format. We use 96-well PCR plates for all of our applications. Set diluted samples aside on ice.
3. Preheat a thermal cycler using the heated lid option to 72 $^{\circ}$ C.
4. Prepare 10 \times reaction buffer (19 μ l 10 \times lysis buffer + 1 μ l RNase inhibitor). Do not vortex, as the lysis buffer contains a detergent (*see Note 17*). Add 1 μ l 10 \times reaction buffer to each 9.5 μ l sample, and the positive and negative controls. All samples should have a final volume of 10.5 μ l. Set samples aside on ice.
5. Add 2 μ l of 3' SMART-Seq[®] CDS Primer IIA (12 μ M) to each sample (*see Note 18*). Pipette up and down using a multichannel pipette to mix. Close with your seal of choice or cap tubes, and spin briefly. Proceed to next step.

6. Incubate the samples at 72 °C in a preheated, hot lid thermal cycler for 3 min.
7. While samples are incubating at 72 °C, prepare the next master mix at room temperature. For each sample, add 4 µl 5× Ultra Low First-Strand Buffer + 1 µl SMART-Seq[®] v4 Oligonucleotide (48 µM) + 0.5 µl RNase Inhibitor (49 U/µl). Be sure to include overage (*see* **Note 19**).
8. Immediately after the 3 min incubation, place samples on ice for 2 min.
9. Preheat the thermal cycler to 42 °C.
10. While samples are on ice, remove the SMARTScribe Reverse Transcriptase (RT) from the freezer and place in a benchtop -20 °C cooler or on ice (*see* **Note 20**). DO NOT VORTEX the reverse transcriptase before adding it to the master mix, simply invert up and down or pipette gently. Add 2 µl of SMARTScribe Reverse Transcriptase (RT) *per reaction* to the master mix. Be sure to use the same amount of samples used in Subheading 3.4, step 7 to determine how much RT to use. After the RT is added to the master mix, the master mix tube can be gently vortexed, and then spun to collect droplets at bottom of tube
11. Add 7.5 µl of master mix with RT to each sample. Pipette up and down using a multichannel pipette to mix. Close with seal of choice or cap tubes and spin briefly. Proceed to next step.
12. Place samples on preheated thermal cycler at 42 °C with a heated lid. Run the following program: 42 °C for 90 min/70 °C for 10 min/4 °C hold. Samples may be stored at 4 °C overnight or the protocol can be followed as described.
cDNA amplification by LD-PCR (see Note 21)
13. Reagents should be on ice except for the SeqAmp DNA polymerase, which should be at -20 °C. Gently vortex each reagent and spin down quickly.
14. Remove the SeqAmp DNA Polymerase from freezer and place in a benchtop -20 °C cooler or on ice. DO NOT VORTEX; simply invert to mix.
15. Combine the following reagents for the amplification master mix on ice in the following order: 25 µl 2× SeqAmp PCR Buffer + 1 µl PCR Primer IIA (12 µM) + 1 µl SeqAmp DNA Polymerase + 3 µl nuclease-free water. Account for number of samples plus overage.
16. Once all reagents have been added to master mix, vortex briefly and spin down to collect droplets to the bottom.
17. Add 30 µl of cDNA amplification master mix to each reaction containing 20 µl of first-strand cDNA product. Pipette up and down using a multichannel pipette to mix. Close with your

Amount of starting total RNA	Amount of starting cells	Suggested number of PCR cycles
10 ng	1,000	7-8
1 ng	100	10-11
100 pg	10	14-15
10 pg	1	17-18

Fig. 4 Suggested PCR cycles based on starting RNA amount. The chart above shows the number of PCR cycles suggested depending on the amount of total RNA or the number of cells the protocol started with

choice of seal or cap tubes and spin briefly. Proceed to next step.

18. The number of cycles of cDNA amplification is determined by the input amount of total RNA (*see Note 22*, Fig. 4).
19. Place in a preheated thermal cycler with a heated lid and run the following program: 95 °C for 1 min/98 °C for 10 s/ 65 °C for 30 s/68 °C for 3 min/return to **steps 2–4** for number of cycles indicated in above chart/72 °C for 10 min/hold at 4 °C forever. At this point, samples may be stored at 4 °C overnight.

Purification of Amplified cDNA using Agencourt AMPure XP beads

20. Thaw AMPure XP beads for >30 min. Make 80% ethanol (*see Note 23*). You will need a magnetic separation device that is compatible with either plate or PCR tube format (*see Note 24*).
21. Add 1 µl of 10× Lysis Buffer to each sample, mix by pipetting up and down.
22. Vortex AMPure XP beads well to obtain a homogenous mixture of beads. Add 50 µl of AMPure XP beads to each sample.
23. Mix by pipetting up and down with a multichannel pipette, until the bead/sample mixture is homogeneous. Close with choice of seal and spin briefly.
24. Incubate for 8 min at room temperature to allow binding between cDNA and beads (*see Note 25*).
25. Briefly spin samples again and place on magnet for ~5 min until liquid appears clear (*see Note 26*).
26. While samples are on magnet, remove the supernatant using a pipette and discard (*see Note 27*).
27. With samples still on magnet, immediately add 200 µl of 80% ethanol to samples and let sit for 30 s. Pipette off supernatant and discard.
28. Repeat ethanol wash (Subheading 3.4, step 27) once more.
29. Seal or cap samples and spin down briefly to collect any remaining liquid. Immediately place samples back on magnet and let

- sit for 30 s (still sealed or capped). Remove seal or caps and remove any remaining ethanol with pipette.
30. Allow samples to sit for ~2 min to dry bead pellet (*see Note 28*).
 31. Once beads are dry, add 17 μ l of Elution Buffer to cover bead pellet. Remove samples from magnet and mix well (*see Note 29*).
 32. Incubate samples for 2 min at room temperature to rehydrate (*see Note 30*).
 33. Quick-spin samples and place back on magnet for 2 min or longer until solution is completely clear (*see Note 31*).
 34. Transfer 15–16 μ l of clear supernatant (containing cDNA) to new plate or tube (*see Note 32*).
 35. You may stop here and store samples at -20 °C indefinitely.
Validation of cDNA using the High Sensitivity (HS) DNA kit and BioAnalyzer from Agilent
 36. If the kit is new, the gel–dye mix must be prepared (*see Note 33*). Allow the gel–dye mix and other kit components to equilibrate to room temperature for 30 min.
 37. Adjust the chip priming station to the correct setting for DNA chips. Note that the silver lever is at a different position than with an RNA chip. The lever should be at the bottom-most position for DNA chips.
 38. Put a new HS DNA chip on the chip priming station and pipette 9 μ l of gel–dye mix into the well marked © corresponding to the third well down on the right side (Fig. 5).
 39. Be sure that the chip plunger on the chip priming station is at 1 ml, lower the lid of the chip priming station, and push the plunger down until the top of the plunger hooks under the silver tab to hold the plunger in place.
 40. Let the chip sit for exactly 60 s and release the plunger by pulling up on the silver tab, allowing the plunger to pop back up and slowly return to the 1 ml position. Let the plunger retract back toward the 1 ml position for 5 s, and then slowly pull the plunger back to the 1 ml position (*see Note 11*).
 41. Once the plunger is returned to the 1 ml position, unlock the chip priming lid from the front of the chip priming station, and add 9 μ l of gel–dye mix to the three remaining “G” wells.
 42. Load 5 μ l of the RNA marker (green-topped tube) to each sample well of the chip, including the ladder position, but excluding the “G” wells that already contain gel–dye mix (*see Note 12*).
 43. Pipette 1 μ l of ladder into ladder well.

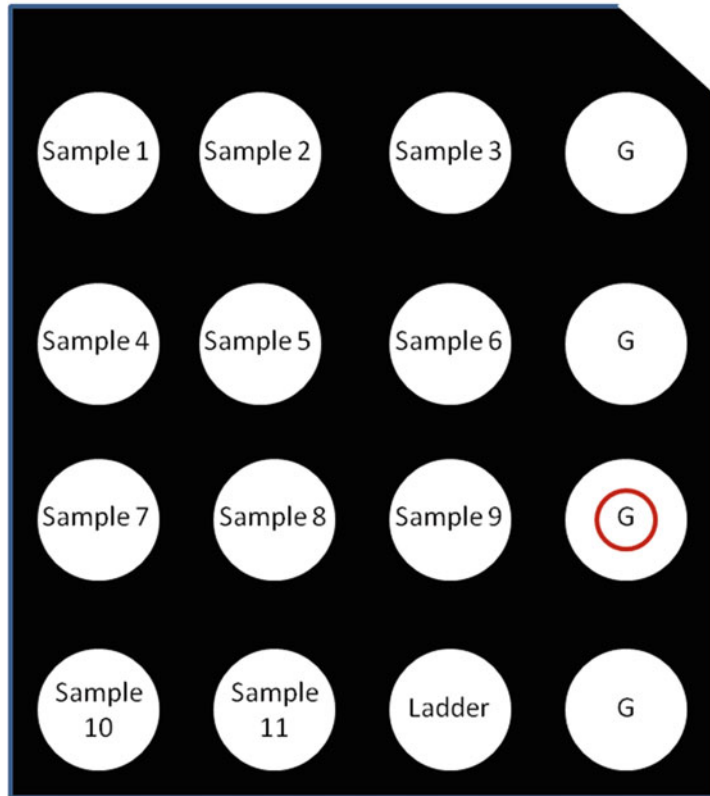


Fig. 5 Loading a High Sensitivity DNA Bioanalyzer chip. Gel-dye mix is added into the well with the circled G. Once the chip is primed, add gel-dye mix to 3 remaining wells marked G. Add marker to all sample and ladder wells. Add samples in chronological order. Add ladder to the well marked “ladder”

44. Pipette 1 μ l of sample into each sample well. If there are fewer samples than sample spaces, add 1 μ l of RNA marker (green-topped tube) to each empty well.
45. Place loaded chip horizontally onto the IKA vortexer and vortex for 1 min at \sim 2200 rpm (*see Note 34*, Fig. 6).
46. Run the chip on the BioAnalyzer within 5 min using the High Sensitivity DNA assay.
47. Analyze the profile. The sample should peak at \sim 2500 bp, and the negative control should be flat between 500 and 9000 bp (Fig. 7).

3.5 Library Preparation Using Illumina Nextera[®] XT DNA Library Kit

The full-length cDNA can be used with the Nextera[®] XT DNA library kit. The Nextera[®] XT protocol can be followed using 1 ng cDNA, and as little as 100–150 pg of amplified cDNA is sufficient.

1. Based on the concentration obtained with the BioAnalyzer, samples are diluted accordingly for measurement on the



Fig. 6 Modified IKA vortexer settings. The IKA vortexer should be set to ~2200 rpm when vortexing a DNA chip to avoid marker carryover from one well to the next. Estimate the middle of the two set points (2000 and 2400) on the vortexer, draw a line and manually set your knob to that speed

Qubit. If the concentration of cDNA is >1 ng/ μ l, dilute the samples at least 1:2 and run on the Qubit High Sensitivity DNA assay. The goal is to use the Qubit measurement for Tagmentation, while taking >2 μ l of sample into the Tagmentation protocol. High Sensitivity DNA standards, High Sensitivity buffer, and dye are required (*see Note 35*).

2. Use the diluted concentrations based on the DNA High Sensitivity Qubit for each sample to determine how much volume of sample is needed to proceed to the Nextera[®] XT protocol, using 1 ng total in 5 μ l. Reactions can be set up in 96-well plate format.

Tagmentation

3. Remove the cDNA, ATM, and TD from -20 °C and thaw on ice. Invert 3–5 times to mix. The NT is stored at room temperature and should be vortexed until all particulates are resuspended.

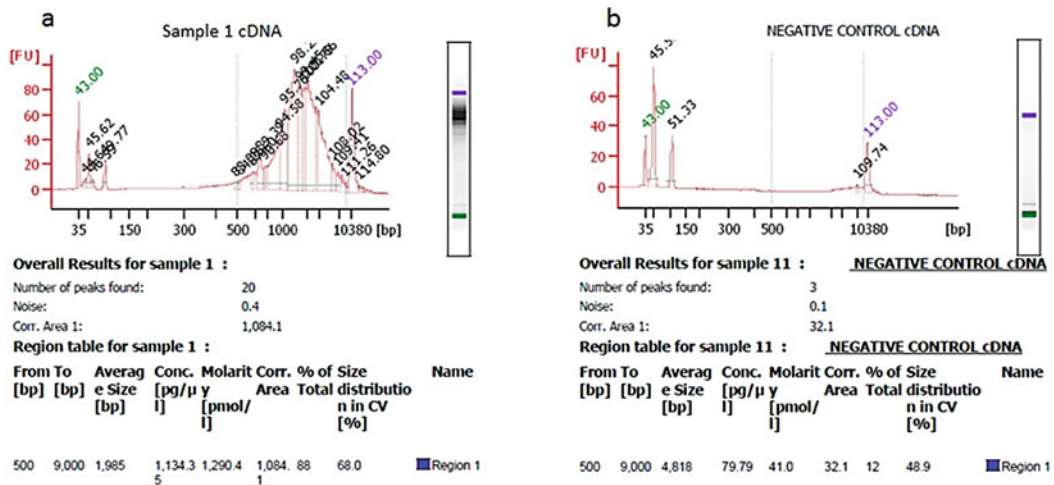


Fig. 7 cDNA profiles on a High Sensitivity DNA Bioanalyzer chip. **(a)** A High Sensitivity DNA BioAnalyzer chip showing a trace of good cDNA from 2 ng total RNA using the SMART-Seq[®] v4 Ultra[®] Low kit and 11 cycles of PCR. cDNA profiles should be between 400 and 10,000 bp peaking at ~2500 bp. The yield of cDNA should be between 3.4 and 17 ng total. The yield is dependent on input amount and type of sample. **(b)** A High Sensitivity DNA BioAnalyzer chip showing the negative control from total RNA using the SMART-Seq[®] v4 Ultra Low[®] kit and 11 cycles of PCR

4. Put the following program into a thermal cycler, choose the heated-lid option, and label the protocol “Tagmentation”:
55 °C for 5 min/10 °C hold.
5. Add 10 μl of TD to each sample well of a 96-well plate. Add 5 μl of the 1 ng sample to plate. Pipette up and down using a multichannel pipette to mix. Seal plate or cap tubes and spin briefly.
6. Add 5 μl ATM to each well. Pipette up and down using a multichannel pipette to mix. Seal plate or cap tubes and spin briefly.
7. Place on preprogrammed thermal cycler and run the “Tagmentation” protocol.
8. Remove plate from cycler and add 5 μl NT to each well. Pipette up and down using a multichannel pipette to mix. Seal or cap tubes and spin briefly.
9. Incubate at room temperature for 5 min.
10. Optional: To assess tagmentation, run 1 μl sample on the BioAnalyzer instrument using the High Sensitivity DNA chip (see **Note 36**).

Amplify Libraries

This step amplifies the tagmented DNA using a limited-cycle PCR program. This step adds the Index 1 (i7) adapters and Index 2 (i5) adapters and sequences required for cluster formation [9].

11. Remove the i5 and i7 adapters from $-20\text{ }^{\circ}\text{C}$ and thaw at room temperature for 20 min. Invert each tube, spin briefly, and set aside (*see Note 37*). Remove the NPM from $-20\text{ }^{\circ}\text{C}$ and thaw on ice.
12. Program the thermal cycler with the Library Amplification Protocol: $72\text{ }^{\circ}\text{C}$ for 3 min/ $95\text{ }^{\circ}\text{C}$ for 30 s/ $95\text{ }^{\circ}\text{C}$ for 10 s/ $55\text{ }^{\circ}\text{C}$ for 30 s/ $72\text{ }^{\circ}\text{C}$ for 30 s/return to **steps 3–5** for 12 cycles/ $72\text{ }^{\circ}\text{C}$ for 5 min/ $10\text{ }^{\circ}\text{C}$ hold.
13. Arrange the i7 adapters in columns 1–12 of the TruSeq Index Plate Fixture. Arrange the i5 adapters in rows A–H of the TruSeq Index Plate Fixture. Place plate or PCR tubes in a PCR rack in open space of Index Plate Fixture. If building less than 12 libraries, it is imperative that the Nextera[®] Low Plex Pooling guidelines from Illumina be consulted (*see Note 38*; Fig. 8).
14. Using a multichannel pipette, add $5\text{ }\mu\text{l}$ of i7 index to each sample column. Replace the caps with new orange caps for each index.



Fig. 8 Nextera[®] XT dual indexing set up. Nextera[®] XT index set up using Illumina's TruSeq Index Plate Fixture. Index 1 (i7) goes across the top of the plate (*orange caps*) while Index 2 (i5) goes vertically down the side (*white caps*). Add i7 indexes down each column and add i5 indexes across each row. If making less than 12 libraries please see the Illumina low plexing guidelines

15. Using a multichannel pipette, add 5 μl of each i5 index to each row with sample. Replace the caps with new white caps for each index.
16. Add 15 μl NPM to each well containing sample. Pipette up and down using a multichannel pipette to mix. Seal plate or cap tubes and spin briefly (*see Note 39*).
17. Place in thermal cycler and run the Library Amplification protocol.
18. This is a safe stopping point where samples can be stored at 4 $^{\circ}\text{C}$ for up to 2 days.

Library Clean up with AMPure XP beads—Illumina recommends switching samples to a midi plate; however, we keep our samples in 96-well PCR plates and pipette up and down to mix.

19. Remove the RSB from -20°C and thaw at room temperature. RSB can be stored at 4 $^{\circ}\text{C}$ after the initial thaw.
20. Remove the AMPure XP beads from 4 $^{\circ}\text{C}$ and let sit for >30 min at room temperature.
21. Prepare 80% ethanol (*see Note 23*).
22. Quick spin the samples from Library Amplification.
23. Determine that there is 50 μl of PCR product (*see Note 40*).
24. Add 30 μl AMPure XP beads to each well. This is $0.6\times$ AMPure XP bead volume (*see Note 41*).
25. Pipette up and down 10–15 times to mix.
26. Incubate at room temperature for 5 min.
27. Place on magnetic stand and let sit for 2 min or until liquid is clear.
28. Remove and discard the supernatant from each well.
29. With the plate on the magnet, add 200 μl of 80% ethanol, incubate for 30 s, remove and discard all supernatant.
30. Repeat Subheading 3.5, step 29 once more for a total of two washes.
31. Seal the plate, quick spin, and place back on magnet. Let sealed plate sit on magnet for 30 s and remove all remaining ethanol.
32. Air dry on magnet for 2–15 min (*see Note 42*).
33. Remove plate from magnetic stand and resuspend in 52.5 μl RSB to each well.
34. Pipette up and down 10–15 times to mix thoroughly.
35. Incubate plate at room temperature for 2 min (*see Note 43*).
36. Place on magnet until liquid is clear, at least 2 min.
37. Remove seal and transfer 50 μl to 1.7 ml eppendorf tube or new PCR plate (*see Note 44*).

38. Safe stopping point. Samples may be stored at $-20\text{ }^{\circ}\text{C}$ indefinitely (*see Note 45*).
- QC libraries on High Sensitivity BioAnalyzer chip*
39. Run $2\text{ }\mu\text{l}$ of library using the High Sensitivity DNA Qubit assay. Reference Qubit QC in Subheading 3.2 being sure to use the DNA High Sensitivity kit. If concentration is greater than $1\text{ ng}/\mu\text{l}$ dilute to $\sim 1\text{ ng}/\mu\text{l}$ to run on the BioAnalyzer.
 40. Run $1\text{ }\mu\text{l}$ of library, diluted or neat, dependent on Qubit concentration, on a BioAnalyzer DNA High Sensitivity Chip. Refer to Subheading 3.4, step 36.
 41. Typical libraries show a broad size distribution of $\sim 250\text{--}1000\text{ bp}$. A wide variety of libraries can be sequenced with fragments as small as 250 bp or as large as 1500 bp (Fig. 9).
 42. Proceed to normalization of samples using the reagents supplied in the Nextera[®] XT kit or pool by hand, based on manual QC values (*see Note 46*). We choose to pool based on QC values from the BioAnalyzer and Qubit, because we like the ability to keep our libraries at full concentration while diluting only a portion of the library for pooling.
 43. To calculate molarity, take the $\text{ng}/\mu\text{l}$ concentration from the Qubit and multiply it by 10^6 . Divide that number by $649 \times$ average bp size from the High Sensitivity BioAnalyzer

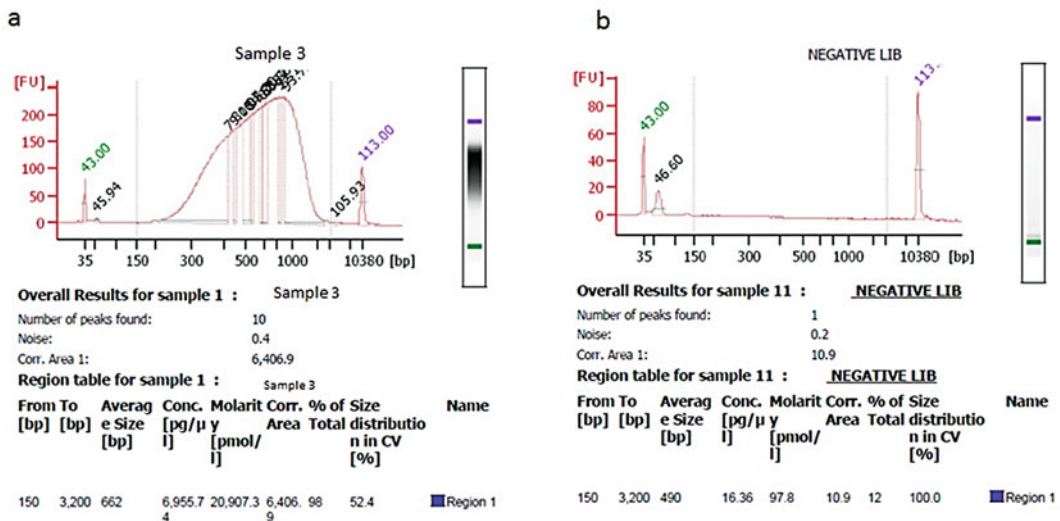


Fig. 9 Successful library profile. (a) Completed library from 3 ng of total RNA through the SMART-seq[®] V4 Ultra[®] Low Input kit followed by Nextera[®] XT library build. $1\text{ }\mu\text{l}$ library on High Sensitivity DNA BioAnalyzer chip. Libraries can show a broad size range from 250 to 1000 bp (or larger). Libraries can successfully be sequenced with a size of up to 1500 bp. (b) The negative control gives no library after completion of the protocol

$$\frac{(ng/\mu l \times 10^6)}{(649 \times \text{average library fragment length})} = \text{Molarity}$$

Fig. 10 Molarity equation. To determine molarity use the equation above. In the ng/μl place put the ng/μl from the Qubit High Sensitivity DNA reading, multiply that by 10⁶. Divide the entire top value by 649 × the average bp fragment for each library given by the High Sensitivity BioAnalyzer DNA chip

Library name	bp	nM	ng/ul
Sample 1	798	12.80	6.63
Sample 2	878	10.58	6.03
Sample 3	738	17.68	8.47
Sample 4	768	16.21	8.08

Fig. 11 Excel sheet molarity calculator. Enter the average bp size from the High Sensitivity DNA BioAnalyzer chip into the bp column. Enter the ng/μl from the High Sensitivity DNA Qubit into the ng/μl column. Enter the formula = [(ng/μl × 10⁶)/(649 × average library fragment length)] into the nM column and the excel sheet will figure out your molarity based on the average bp size of your library and the concentration in ng/μl

measurement. To obtain the average bp size, set the regions of interest on the BioAnalyzer between 250 bp and 1000 bp or 1500 bp as the protocol above suggests (Fig. 10).

44. The above formula can be placed into an excel sheet (in yellow box) and Qubit concentrations (ng/μl column) and average basepair size (bp column) inputted, and the formula will determine the nM column automatically. Be sure to use double parentheses in your equation (yellow box; Fig. 11).
45. We recommend creating a spreadsheet in Microsoft Excel to determine sample and diluent volume to achieve a specific molarity. In our example, we are diluting libraries to 10 nM in 10 mM Tris-HCl in a final volume of 10 μl. This is a simple C1V1 = C2V2 calculation put into an excel sheet.
 - (a) C1 = nM concentration of sample,
 - (b) V1 = × (amount of library to add)
 - (c) C2 = 10 nM (final concentration of library)
 - (d) V2 = 10 μl (final volume of diluted sample)

In the excel sheet, the column labeled “vol of library to 10 nM” is V1 (V1 = C2×V2/C1) so V1 = (10 nM × 10 μl)/nM of library. To determine the volume of diluent needed [column “vol of 10mM Tris-HCl pH 8.0 (10 μl)”], take the total volume (V2 or 10 μl) and subtract V1 (volume of library) (Fig. 12).

Library Name	nM	Vol of library to 10nM	Vol of 10mM Tris-Cl pH 8.0 (10ul)
Sample 1	12.59	7.94	2.03
Sample 2	10.41	9.61	.39
Sample 3	17.39	5.75	4.25
Sample 4	15.94	6.27	3.73

Fig. 12 Library dilution spreadsheet calculations. Type the nM concentration of your libraries based on the formula from Fig. 10 into the nM column. Input the formula = $C2 \times V2 / C1$ into the column labeled “Vol of library to 10 nM” so = $(10 \text{ nM} \times 10 \mu\text{l}) / \text{nM of library (column 1)}$. The 3rd column is 10 μl so subtract the 2nd column from 10 μl to determine how much 10 mM Tris–HCl should be added to each diluted sample to obtain a 10 nM dilution of each sample. You can dilute your samples to any concentration, keeping in mind your least concentrated sample

46. Once each sample is diluted, pool equal volumes of each sample into a 1.7 ml tube.
47. Once libraries are pooled, they are ready to be turned over to the Genomics Core Facility for sequencing on any Illumina platform or stored at $-20 \text{ }^\circ\text{C}$ indefinitely.

4 Notes

1. A spectrophotometer will also work because only a preliminary concentration is needed to determine which Qubit assay to use.
2. Nextera[®] XT, not Nextera kit. Be sure to purchase the correct kit.
3. The 260/230 ratio of ~ 2.0 is generally accepted as “pure” for RNA and 2.0–2.2 is acceptable. If the ratio is lower, it may indicate contamination that absorbs at 230 nm such as EDTA, carbohydrates, or phenol.
4. Note that to measure 1 ng/ μl with the BR assay, 20 μl of sample will need to be measured. The range of detection is indicative of the amount of sample used. With a sample at 1 $\mu\text{g}/\mu\text{l}$ on the NanoDrop, 1 μl can be used with the BR assay to get an accurate concentration.
5. To measure 250 pg/ μl , 20 μl of sample will need to be added. To measure 100 ng/ μl , only 1 μl of sample will need to be added. It is possible to determine concentration using only 2 μl of sample.
6. Qubit tubes must be used for this assay. No other 0.5 ml tubes will give accurate results. Be sure to order the special Qubit tubes when ordering the Qubit assay.
7. Do not write on the sides of the Qubit tubes as this may interfere with the readings.

8. The Qubit assay is light sensitive. If the buffer + dye + sample mixture will sit for longer than 2 min before reading on the Qubit machine, place samples in a dark place.
9. Please reference the correct Qubit protocol for the type of assay you are doing. In the case of the BR RNA assay: https://tools.thermofisher.com/content/sfs/manuals/Qubit_RNA_BR_Assay_UG.pdf

After Subheading 3.5 skip to page 6 step 4.1 to let the Qubit determine the concentration of the sample for you.

10. RNA Pico Guide http://www.agilent.com/cs/library/usermanuals/Public/G2938-90049_RNA6000Pico_QSG.pdf
11. We like to see the plunger bounce back to the 0.6 ml or 0.7 ml mark and then slowly retract further toward the 1 ml mark of the plunger. If the plunger does not bounce back to the 0.6 ml mark immediately, this can indicate that there is a hole in the gasket of the plunger. Full instructions for replacing the gasket can be found in the BioAnalyzer kit documents.
12. Switch tips between each well of the chip when adding marker. Also go straight down into the well with the tip, not at an angle and don't push through to cause a bubble in the liquid.
13. Figure 3 shows a good total RNA trace on the Pico chip from the BioAnalyzer. The 18 s and 28 s peaks are sharp, the FU (fluorescent unit) measurement on the Y-axis is high lending to believability that it is a good run and RNA is concentrated. The RIN score is 9.6, which is above our cutoff of 7.5 for this protocol.
14. We work in 96-well plate format. We pipette samples up and down to mix well before sealing plate. All incubation steps require a seal for the plate before going on the thermal cycler. We prefer an adhesive plate seal for all of our incubation procedures.
15. We have been successful in this protocol down to 300 pg of total RNA basing total RNA concentration off of Pico BioAnalyzer data.
16. The positive control is supplied at 1 $\mu\text{g}/\mu\text{l}$, and we dilute it to 1 $\text{ng}/\mu\text{l}$ using nuclease-free water.
17. With more than 18 samples, additional 10 \times reaction buffer will be needed. Because the solution is foamy, expect to lose two reaction volumes per master mix.
18. If 17 or more PCR cycles are being performed, use 1 μl of 3' SMART-Seq[®] CDS Primer II A. Keep the final volume at 12.5 μl by increasing the volume of RNA/cells in validated media to 10.5 μl , either by adding additional nuclease-free water or by increasing sample volume. Keep the volume of 10 \times Reaction Buffer at 1 μl regardless of the number of PCR cycles.
19. Make at least 10% excess for each master mix.

20. We use a benchtop -20°C cooler. When not in use, it is stored at -20°C at all times.
21. PCR Primer II A amplifies cDNA from the SMART sequences introduced by 3' SMART-Seq[®] CDS Primer II A and the SMART-Seq[®] v4 Oligonucleotide.
22. Because 3 ng of total RNA was used, cDNA amplification at 10 cycles was performed.
23. Make fresh 80% ethanol daily. Make enough for that day, including overage. Each sample requires 400 μl of 80% ethanol.
24. We use the DynaMag-96 side magnet from Thermo Fisher for plate set up. It has a 13th column for sample mixing by moving the plate back and forth from right to left on the magnet. The DynaMag-PCR magnet works well for PCR tube set up.
25. You can incubate for as long as 15 min. The lower the amount of total starting RNA, the longer the incubation time. If cDNA is left unbound to beads in the supernatant, it can be discarded in the following two steps, which could affect yield.
26. If supernatant appears cloudy after 5 min allow to sit on magnet for another 5 min. Sometimes a brown halo around the bead pellet will be seen. This is not unusual.
27. Avoid the bead/halo pellet, so sample will not inadvertently be sucked up and discarded. Sucking up beads at this point could cause a lower yield than expected.
28. When samples first begin to dry, the bead pellet will appear shiny. Once the samples have dried, the pellet should be matte in appearance, but not cracked. In a humid climate, it may take a little longer to dry the pellet. In a dry climate, it usually takes 2–3 min to dry the pellet. We recommend checking the level of dryness every 30 s, because overdrying the pellet to the point of cracks will reduce cDNA yield. If cracks appear in the samples, add Elution Buffer to the pellets and let them sit for 5–15 min. In addition to overdrying the pellet, leftover ethanol will also reduce cDNA yield.
29. Samples can be mixed by pipetting up and down, being sure to elute all beads off sides of tube or vortexing. Spin down samples after mixing.
30. Samples should be homogeneous when mixed with Elution Buffer. To ensure adequate rehydration, samples can be incubated at room temperature longer than 5 min.
31. Incubation for 2 min is the minimum amount of time recommended, as it is important that all beads bind to the magnet.
32. We transfer to a 1.7 ml tube labeled with each sample name so that we can continue on to QC of cDNA by BioAnalyzer. It is easier to keep track of samples on the BioAnalyzer chip when samples are not in plate format.

33. If the HS DNA BioAnalyzer kit is new, thaw the kit to room temperature for 30 min. Vortex the High Sensitivity dye (blue cap) for 10 s and spin down briefly. Pipette 15 μ l of dye mixture into the High Sensitivity Gel Matrix (red cap). Cap the High Sensitivity Gel Matrix and vortex for 10 s and quick spin. Transfer entire volume of gel-dye mix to spin column provided in kit. Do not touch pipette tip to spin column matrix. Spin at $2240 \times g$ ($\pm 20\%$) for 10 min. Discard filter and label gel-dye matrix with date. Gel-dye mix is good for 6 weeks and sufficient for five chips. Store at 4 °C and protected from light, preferably in the original box with the remaining kit components.
34. Vortexing DNA chips at 2400 rpm can sometimes lead to marker carryover. Therefore, we recommend vortexing at 2200 rpm to eliminate the long running of the upper marker into the next sample. It is helpful to mark this speed on the vortexer by drawing a line between 2000 rpm and 2400 rpm.
35. For good concentration readings, samples can be diluted in a ratio of 1:2 before measurement using the Qubit instrument. The Qubit is recommended in the Nextra XT protocol to measure input DNA concentrations. We recommend using $>1 \mu$ l and preferably, $> 2 \mu$ l in the tagmentation protocol.
36. We do not assess quality at the tagmentation step. If problems arise with the library protocol, we recommend diluting the original cDNA samples and rerunning the tagmentation protocol. With very low cDNA yields, we recommend checking tagmentation at this step.
37. A benchtop centrifuge can be used to spin down adapters by first placing a 1.7 ml eppendorf tubes in the centrifuge, and then placing the adapter tubes in the 1.7 ml tubes. Spin as normal.
38. Be cautious when only doing a few samples. The i7 indexes need to be different when multiplexing which may be counter-intuitive if a vertical method is used for working in plate format. Illumina provides a low-plex pooling document that should be addressed when doing less than 12 samples. **This is very important.**
39. Enough NPM can be aliquoted into each well of an 8-strip tube to enable use of a multichannel pipette for pipetting that reagent into samples. We account for at least 10% overage in this situation.
40. The ratio of PCR product to AMPure XP beads needs to be 3:2. For example, 50 μ l PCR product to 30 μ l AMPureXP beads.
41. The Illumina protocol recommends transferring samples to a midi plate before proceeding with AMPure XP bead clean

- up. We have found that cleanup in the PCR plate works well and pipetting up and down to mix thoroughly is sufficient.
42. Illumina recommends drying for 15 min, but this length of time will cause beads to crack. We dry only until the beads look matte, but not cracked.
 43. We let our plates incubate at room temperature for 5 min, especially if beads are drier than anticipated.
 44. We find QC is easier when libraries are stored in individual eppendorf tubes, especially if doing this protocol on more than one project. With a high throughput QC machine, it may be easier to leave the samples in plate format.
 45. According to the Illumina protocol, samples can be stored for up to 7 days. Libraries will be stable at -20°C indefinitely.
 46. Please refer to the Nextera[®] XT DNA Library Prep Reference Guide for normalization of libraries. If the final concentration of the library is less than 15 nM, normalization is NOT recommended. Normalization will yield a single-stranded library that is pooled in equal volumes in a 1.7 ml tube, prior to loading on the sequencer. If sending samples to a Core Sequencing Facility, let the staff know that the samples were pooled in this way, as it changes the clustering protocol downstream in sequencing.

References

1. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW (2016) Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet* 17:257–271
2. Hrdlickova R, Toloue M, Tian B (2017) RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA* 8
3. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10:57–63
4. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublotte JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, Trombetta JJ, Gennert D, Gnirke A, Goren A, Hacohen N, Levin JZ, Park H, Regev A (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498:236–240
5. Tirosh I, Izar B, Prakadan SM, Wadsworth MH 2nd, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, Fallahi-Sichani M, Dutton-Regester K, Lin JR, Cohen O, Shah P, Lu D, Genshaft AS, Hughes TK, Ziegler CG, Kazer SW, Gaillard A, Kolb KE, Villani AC, Johannessen CM, Andreev AY, Van Allen EM, Bertagnolli M, Sorger PK, Sullivan RJ, Flaherty KT, Frederick DT, Jane-Valbuena J, Yoon CH, Rozenblatt-Rosen O, Shalek AK, Regev A, Garraway LA (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352:189–196
6. Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, Chen P, Gertner RS, Gaublotte JT, Yosef N, Schwartz S, Fowler B, Weaver S, Wang J, Wang X, Ding R, Raychowdhury R, Friedman N, Hacohen N, Park H, May AP, Regev A (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 510:363–369
7. Trombetta JJ, Gennert D, Lu D, Satija R, Shalek AK, Regev A (2014) Preparation of single-cell RNA-seq libraries for next generation sequencing. *Curr Protoc Mol Biol* 107:4.22.21–4.22.17
8. Clontech Laboratories, Inc. (2017) Clontech SMART-seq ultra low input RNA kit for sequencing. <http://www.clontech.com>. Accessed 7 Jan 2017
9. Illumina, (2017) Nextera XT DNA library preparation kit. <https://www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/nextera-xt-dna.html>. Accessed 7 Jan 2017

Chapter 11

Using Fluidigm C1 to Generate Single-Cell Full-Length cDNA Libraries for mRNA Sequencing

Robert Durruthy-Durruthy and Manisha Ray

Abstract

Single-cell RNA sequencing has evolved into a benchmark application to study cellular heterogeneity, advancing our understanding of cellular differentiation, disease progression, and gene regulation in a multitude of research areas. The generation of high-quality cDNA, an important step in the experimental workflow when generating sequence-ready libraries, is critical to maximizing data quality. Here we describe a strategy that uses a microfluidic device (i.e., the C1™ IFC) to synthesize full-length cDNA from single cells in a fully automated, nanoliter-scale format. The device also facilitates confirmation of the presence of a *single, viable* cell and recording of phenotypic information, quality control measures that are crucial for streamlining downstream data processing and enhancing overall data validity.

Key words Single-cell RNA-seq, Gene expression profiling, Full-length cDNA, Single-cell transcriptomics, Cell characterization

1 Introduction

Since its introduction in 2009 [1], single-cell RNA sequencing has become the method of choice to study transcriptional heterogeneity of tissues and has contributed to new biological insights in a number of research areas, including stem cell biology [2], neurobiology [3], and immunology [4]. To date, more than a dozen techniques have been developed to generate single-cell libraries for next-generation sequencing, each with unique sets of advantages and disadvantages [5]. Despite considerable progress in the development of new, robust protocols, all methods available today suffer from substantial technical (*nonbiological*) variability due to the minute amounts of starting material, hampering downstream data processing and information extraction. Here we present a strategy to generate full-length cDNA from individual cells using a microfluidic device (i.e., integrated fluidic circuit, or IFC from Fluidigm®). Nanoliter reaction volumes, a fully automated workflow, and the ability to visually confirm the presence of a *single*,

viable cell combine to minimize technical noise and maximize overall data. The following protocol describes the generation of full-length cDNA from single cells and outlines critical steps including IFC priming, cell loading, automated lyses, reverse transcription, cDNA synthesis, and cDNA harvesting.

2 Materials

2.1 The C1™ Reagent Kit for mRNA Seq (Fluidigm®) Consists of Module 1 and Module 2

- **Module 1:** Cell Wash Buffer, Suspension Reagent, C1 Blocking Reagent.
- **Module 2:** Loading Reagent, C1 Preloading Reagent, C1 DNA Dilution Reagent, C1 Harvest Reagent.
 - SMARTer® Ultra® Low RNA Kit for the Fluidigm C1 System, 10 IFCs (Takara Bio, Inc.®)

2.2 Box 1, Box 2, Advantage® 2 PCR Kit

1. C1 IFC for mRNA Seq (Fluidigm, select 5–10 µm, 10–17 µm, or 17–25 µm).
2. C1 DNA Dilution Reagent (Fluidigm).
3. Agilent® High Sensitivity DNA chips and reagents (Agilent Technologies).
4. MicroAmp® Clear Adhesive Film (Thermo Fisher Scientific).
5. 96-well PCR plates.
6. C1 system.
7. Centrifuges (for microcentrifuge tubes and 96-well plates).
8. Vortexer.
9. 2100 Bioanalyzer® (Agilent).
10. Thermal cycler.
11. Magnetic stand for PCR tubes.
12. Fluorometer (for PicoGreen® assay).
13. LIVE/DEAD® Viability/Cytotoxicity Kit, for mammalian cells (Thermo Fisher Scientific).
14. ArrayControl™ RNA Spikes (Thermo Fisher Scientific).
15. The RNA Storage Solution (Thermo Fisher Scientific).
16. RNeasy® Plus Micro Kit (Qiagen®).
17. 14.3 M β-mercaptoethanol for 2 M dithiothreitol.
18. QIAshredder™ disposable cell lysate homogenizers (Qiagen).
19. INCYTO C-Chip™ Disposable Hemocytometer (Neubauer Improved).
20. Biocontainment hood.
21. Imaging equipment compatible with C1 IFCs.

22. Nextera[®] XT DNA Sample Preparation Kit (Illumina[®]).
23. Nextera XT DNA Library Preparation Index Kits (Illumina).
24. Quant-IT[™] PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific).
25. Agencourt AMPure[®] XP (Beckman Coulter).
26. Ethanol.

3 Methods

3.1 Prepare Reagent Mixes

3.1.1 (Optional) RNA Spikes Mix

RNA spikes serve as positive controls for thermal cycling of the C1 system independent of cell capture. Although this control is not required, we highly recommend it (*see Note 1*).

1. Thaw the ArrayControl RNA Spikes; remove spikes no. 1, no. 4, and no. 7 from the box.
2. Pipet the following in the three tubes:
 - **Tube A:** 13.5 μL of RNA Storage Solution and 1.5 μL of no. 7 RNA spikes.
 - **Tube B:** 12.0 μL of RNA Storage Solution and 1.5 μL of no. 4 RNA spikes.
 - **Tube C:** 148.5 μL of RNA Storage Solution and 1.5 μL of no. 1 RNA spikes.
3. Vortex Tube A for 3 s and centrifuge to collect contents. Pipet 1.5 μL from Tube A into Tube B. Discard Tube A.
4. Vortex Tube B for 3 s and centrifuge to collect contents. Pipet 1.5 μL from Tube B into Tube C. Discard Tube B.
5. Vortex Tube C for 3 s and centrifuge to collect contents. Tube C is the concentrated RNA standard (RNA Spikes mix), which can be aliquoted in 1.25 μL volumes and stored at $-80\text{ }^{\circ}\text{C}$ for future use (*see Note 2*). One tube is used per C1 run.
6. Dilute the RNA Spikes mix for the Lysis final mix 100 \times by combining 99 μL of C1 Loading Reagent (Fluidigm) with 1.0 μL of RNA Spikes mix (*see Note 3*).
7. Vortex the diluted RNA Spikes mix for 3 s and centrifuge to collect contents.

3.1.2 Lysis Mix—Mix a (See **Note 4**)

1. Mix the following reagents in a tube labeled A (total volume 20 μL):
 - 1.0 μL of diluted RNA Spikes mix [or 1.0 μL of C1 DNA Loading Reagent (Fluidigm) if RNA Spikes mix is not being used].
 - 0.5 μL of RNase Inhibitor (Clontech SMARTer kit).

- 7.0 μL of 3' SMART CDS Primer IIA (Clontech SMARTer kit).
 - 11.5 μL of Dilution Buffer (Clontech SMARTer kit). Do not vortex.
2. Pipet the cell lysis mix up and down a few times to mix. Keep on ice until use.

3.1.3 Reverse Transcription (RT) Mix—Mix B

1. Mix the following reagents in a tube labeled B to create the RT reaction mix (total volume 32 μL):
 - 1.2 μL of C1 Loading Reagent (Fluidigm).
 - 11.2 μL of 5 \times First-Strand Buffer (Clontech SMARTer kit).
 - 1.4 μL of dithiothreitol.
 - 5.6 μL of dNTP Mix (dATP, dCTP, dGTP, and dTTP, each at 10 mM).
 - 5.6 μL of SMARTer IIA Oligonucleotide (Clontech).
 - 1.4 μL of RNase Inhibitor (Clontech).
 - 5.6 μL of SMARTScribe Reverse Transcriptase (Clontech SMARTer kit).
2. Vortex the RT reaction mix for 3 s and centrifuge briefly to collect contents. Keep on ice until ready to use.

3.1.4 PCR Mix—Mix C (See Note 5)

1. Mix the following reagents in a tube labeled C to create the PCR mix (total volume 90 μL):
 - 63.5 μL of PCR-Grade Water (Advantage 2 PCR Kit).
 - 10.0 μL of 10 \times Advantage 2 PCR Buffer [not short amplicon (SA)] (Advantage 2 PCR Kit).
 - 4.0 μL of 50 \times dNTP Mix (Advantage 2 PCR Kit).
 - 4.0 μL of IS PCR Primer (Clontech SMARTer Kit).
 - 4.0 μL of 50 \times Advantage 2 Polymerase Mix (Advantage 2 PCR Kit).
 - 4.5 μL of C1 Loading Reagent (Fluidigm).
2. Vortex the PCR final mix for 3 s and centrifuge to collect contents before use. Keep on ice until ready to use.

3.2 Prime the IFC

When pipetting into the C1 IFC, always stop at the first stop on the pipette to avoid creating bubbles in the inlets. If a bubble is introduced, ensure that it floats to the top of the well. Vortex and then centrifuge all reagents before pipetting into the IFC.

1. Add 200 μL of C1 Harvest Reagent from a 4 mL bottle into each of the accumulators marked with red circles (Fig. 1).

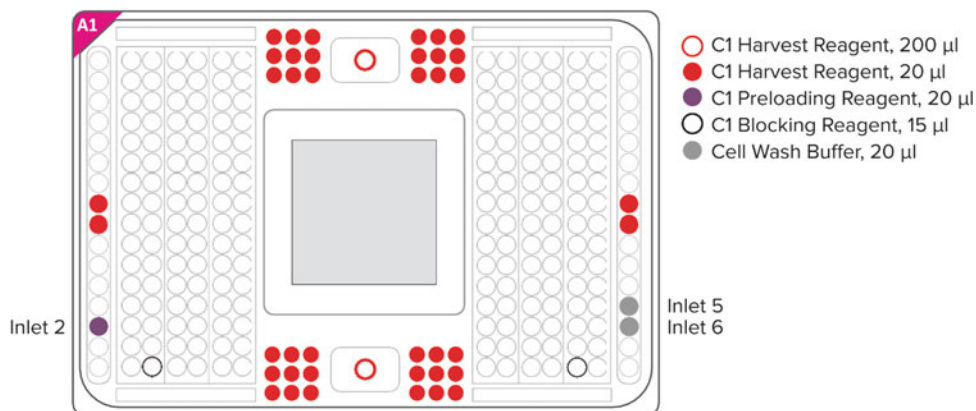


Fig. 1 C1 IFC priming pipetting map. Schematically shown is the C1 IFC. Inlets that are filled with reagents for priming the IFC are color-coded

2. Pipet 20 μL of C1 Harvest Reagent into each inlet marked with solid red circles on each side of the accumulators (36 total).
3. Pipet 20 μL of C1 Harvest Reagent into each of the two inlets marked with solid red circles in the middle of the outside columns of inlets on each side of the IFC. These wells are marked on the bottom of the IFC with a notch to ensure they are easily located.
4. Pipet 20 μL of C1 Preloading Reagent into inlet 2, marked with a purple dot.
5. Pipet 15 μL of C1 Blocking Reagent into the cell inlet and outlet marked with white dots.
6. Pipet 20 μL of Cell Wash Buffer into inlets 5 and 6, marked with dark gray dots.
7. Peel off white tape on bottom of IFC.
8. Place the IFC into the C1 system. Run the **mRNA Seq: Prime** (1771 \times /1772 \times /1773 \times) script. Priming takes approximately 10 min. When the Prime script has finished, tap **EJECT** to remove the primed IFC from the instrument (*see Note 6*).

3.3 Prepare Cells

3.3.1 Prepare LIVE/DEAD Cell Staining Solution (Optional)

The optional live/dead cell-staining step uses the LIVE/DEAD Viability/Cytotoxicity Kit, which tests the viability of a cell based on the integrity of the cell membrane. This test contains two chemical dyes. The first dye is green fluorescent calcein AM, which stains live cells. This dye is cell-permeable and tests for active esterase activity in live cells. The second dye is red fluorescent ethidium homodimer-1, which stains cells only if the integrity of the cell membrane has been lost (*see Note 7*).

1. Vortex dyes for 10 s, and then centrifuge before pipetting.
2. Prepare the LIVE/DEAD staining solution by combining reagents in this order (total volume ~ 1253.13 μL):
 - (a) 1250 μL of Cell Wash Buffer (Fluidigm).
 - (b) 2.5 μL of ethidium homodimer-1 (Thermo Fisher Scientific).
 - (c) 0.625 μL of Calcein AM (Thermo Fisher Scientific).
3. Vortex the LIVE/DEAD staining solution well before pipetting into the IFC.

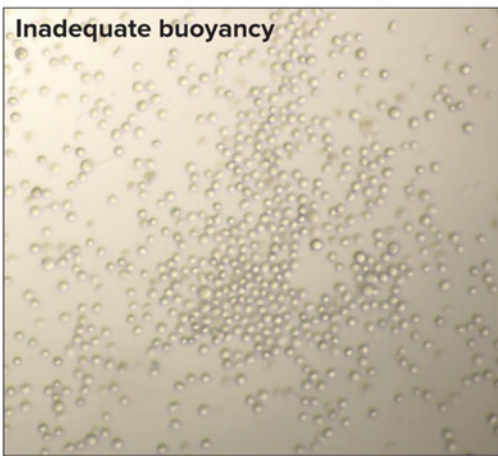
3.3.2 Prepare the Cell Mix (See Note 8)

1. Prepare the cell mix while priming the IFC.
2. Prepare a cell suspension in native medium of 66,000–333,000 cells/mL (Fig. 2). The recommended concentration range ensures that a total of 200–1000 cells are loaded into the IFC. You can prepare a cell suspension with a minimum concentration of 66,000 cells/mL, but fewer cells will be loaded and captured in the IFC. Preparing a cell suspension of

A

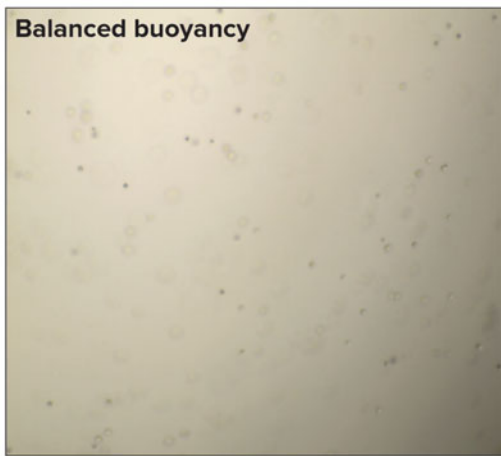
Ratio (%)	50	55	60	65	70	75	80	85	90	100
Cells (μl)	5	5.5	6	6.5	7	7.5	8	8.5	9	10
Suspension Reagent (μl)	5	4.5	4	3.5	3	2.5	2	1.5	1	0
Total volume/well (μl)	10	10	10	10	10	10	10	10	10	10

B



Cells settled on bottom of well

C



Cells suspended through volume of well

Fig. 2 Buoyancy assessment of cells. (a) Titration series of Cell/Suspension Reagent mix across different ratios ranging from 50 to 100%. (b) Representative bright field image of cells that are settled on the bottom of the well indicating incorrect cell/Suspension Reagent ratio (inadequate buoyancy). (c) Representative image of cells that are evenly dispersed throughout the column of the well indicating correct cell/Suspension Reagent ratio (balanced buoyancy)

>333,000 cells/mL may clog the fluidic channels. Suspend the cells in a final volume of 0.5–1 mL to ensure enough cells are available for the IFC and tube controls (*see Note 9*).

3. Prepare the cell mix by combining cells with Suspension Reagent at a ratio of 3:2. For example (total volume 100 μL): 60 μL of cells (concentrated at 166–250/ μL) + 40 μL of Suspension Reagent. The volume of cell mix may be scaled depending on the volume of cells available. A minimum volume of 6 μL of cell mix is necessary for the IFC.
4. Set a P200 pipette to 60 μL , and then pipet the cell mix up and down 5–10 times to mix, depending on whether the cells tend to clump. Do not vortex the cell mix. Avoid bubbles when mixing.

3.4 Load Cells

1. Use a pipette and tip to remove blocking solutions from the cell inlet and outlet marked with teal and white dots, as shown in Fig. 3.
2. Set a P200 pipette to 60 μL , and then pipet the cell mix up and down 5–10 times to mix, depending on whether the cells tend to clump. Do not vortex the cell mix. Avoid bubbles when mixing.
3. Pipet 6 μL of the cell mix into the cell inlet marked with the teal dot. You may pipet up to 20 μL of cell mix, but only 6 μL will enter the IFC.
4. Perform one of these tasks:
 - Staining cells: Vortex the LIVE/DEAD staining solution well, and then pipet 20 μL of the solution into inlet 1, marked with a pink dot.

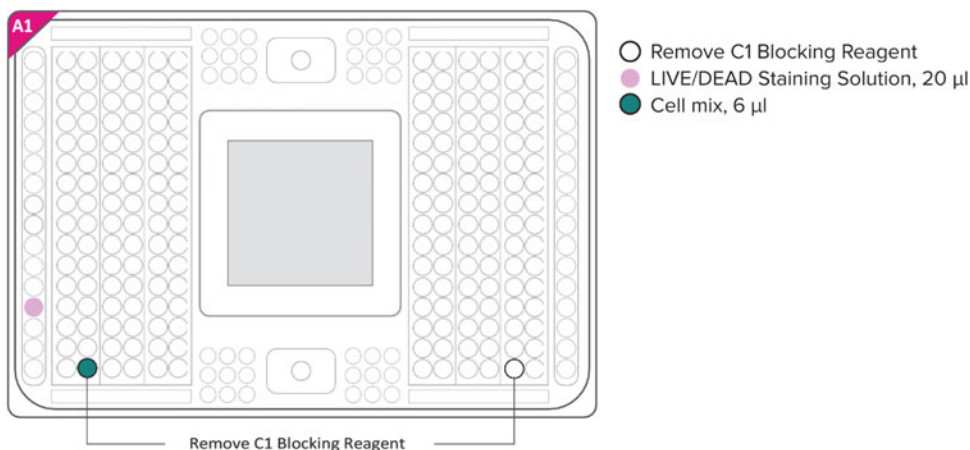


Fig. 3 C1 IFC loading pipetting map. The C1 IFC is schematically shown. Inlets that are filled (or emptied) with reagents for loading the cells on the IFC are color-coded

- Not staining cells: Pipet 20 μL of Cell Wash Buffer into inlet 1, marked with a pink dot (*see Note 10*).
5. Place the IFC into the C1. Run the **mRNA Seq: Cell Load (1771 \times /1772 \times /1773 \times)** or **mRNA Seq: Cell Load & Stain (1771 \times /1772 \times /1773 \times)** script.
 6. When the script has finished, tap **EJECT** to remove the IFC from the C1 system.

3.5 Start the Tube

Control: Lysis and Reverse Transcription (Optional)

Large numbers (e.g., hundreds) of cells in the tube control may inhibit the reaction chemistry. To ensure reliable results, we recommend extraction and purification of total RNA using the Qiagen RNeasy Plus Micro Kit (Qiagen, PN 74034), as described in Protocol A prior to performing the tube controls (*see Note 11*).

3.5.1 Protocol a: Tube Controls with Purified RNA (See Note 12)

1. Dilute cells in media for a final concentration of 100–200 cells/ μL .
2. From this step on, follow the manufacturer's instructions in the Qiagen RNeasy Plus Micro Kit for RNA isolation. Use 20 μL cells in media from the previous step.
3. Proceed to "Perform the Tube Control Reactions" on page 12 of the user's guide, using the purified RNA as the prepared cells in the lysis reaction (*see Note 13*).

3.5.2 Protocol B: Tube Controls with Whole Cells

1. Pellet remaining cells. While speeds and durations of centrifugation may vary, we suggest centrifuging cells at $300 \times g$ for 5 min.
2. Remove the buffer from the pellet by gently pipetting out the supernatant media without disturbing the cell pellet.
3. Resuspend cells in 1 mL Cell Wash Buffer by pipetting up and down at least five times.
4. Pellet cells again and remove supernatant.
5. Wash a second time by resuspending in 1 mL of Cell Wash Buffer by pipetting up and down at least five times.
6. Pellet cells again and remove supernatant.
7. Resuspend cells in Cell Wash Buffer to approximately 90% original volume to keep original concentration, assuming a 10% loss.
8. Dilute your cell suspension to 100–200 cells/ μL using Cell Wash Buffer (*see Note 9*).

3.5.3 Perform the Tube Control Reactions

Prepare cell lysis mix for the positive control by combining 1.0 μL prepared cells and 2.0 μL Lysis final mix. Include a no template control (NTC) by substituting 1.0 μL of Cell Wash Buffer in place of the cells.

1. In a thermal cycler, run a Lysis program corresponding to 72 °C for 3 min, 4 °C for 10 min, 25 °C for 1 min, and then hold at 4 °C.
2. Combine RT final mix (4.0 μL) with lysis thermal products from **step 1** (3.0 μL) for a total volume of 7.0 μL .
3. Vortex the tube controls for 3 s and centrifuge to collect contents.
4. In a thermal cycler, run the following program for reverse transcription: 42 °C for 90 min, 70 °C for 10 min, and then hold at 4 °C (*see Note 14*).
5. When the thermal cycle program is completed, prepare the following reactions for a positive control (Tube 1) and NTC (Tube 2): 9.0 μL PCR mix C, and 1.0 μL RT reaction.
6. Place tubes in a thermal cycler and run the following program: 95 °C for 1 min, 5 cycles of 95 °C for 20 s, 58 °C for 4 min, 68 °C for 6 min, 9 cycles of 95 °C for 20 s, 64 °C for 30 s, 68 °C for 6 min, 7 cycles of 95 °C for 30 s, 64 °C for 30 s, 68 °C for 7 min, 1 cycle of 72 °C for 10 min, and then hold at 4 °C.
7. Transfer prepared material to a post-PCR room.
8. Vortex the prepared products for 3 s and centrifuge to collect contents.
9. Dilute the PCR product by combining 45.0 μL C1 DNA Dilution Reagent and 1.0 μL PCR product.

3.6 Image Cells

Cells may be imaged on a microscope compatible with the C1 IFC (*see Note 15*). Guidelines for the selection of a microscope are outlined in Minimum Specifications for Imaging Cells in Fluidigm Integrated Fluidic Circuits.

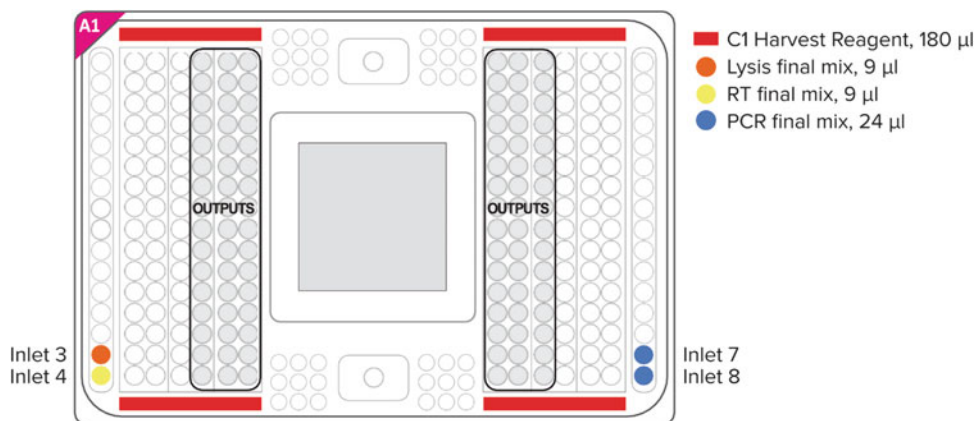


Fig. 4 C1 IFC Lysis, RT and PCR pipetting map. Schematically shown is the C1 IFC. Inlets that are filled with reagents for performing cell lysis, reverse transcription, and PCR on the IFC are color-coded

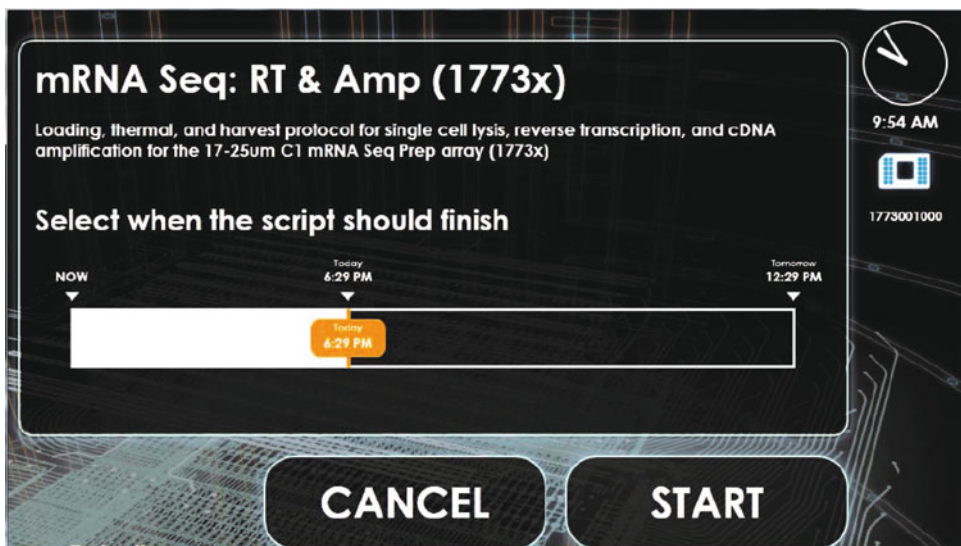


Fig. 5 C1 Display. Screen capture of the C1 display after selecting the PCR script. The user can select the time to harvest the PCR product by sliding the orange button

3.7 Run Lysis, Reverse Transcription, and PCR on the C1 System

1. Pipet 180 μL of C1 Harvest Reagent into the four reservoirs marked with large solid red rectangles in Fig. 4.
2. Pipet 9 μL of lysis Mix A in inlet 3, marked with an orange dot.
3. Pipet 9 μL of RT Mix B in inlet 4, marked with a yellow dot.
4. Pipet 24 μL of PCR Mix C in inlets 7 and 8, marked with blue dots.
5. Place the IFC into the C1 system and run the **mRNA Seq: RT & Amp (1771 \times /1772 \times /1773 \times)** script (Fig. 5, see Note 16).

This protocol can be programmed to harvest at a convenient time. Slide the orange box (end time) to the desired time. For example, the harvest function could be programmed for the next morning. To abort the harvest, tap **ABORT**. The IFC will no longer be usable. Start a new experiment with a new IFC.

The PCR (1771 \times /1772 \times /1773 \times) script contains the thermal cycling protocols shown in Table 1.

3.8 Harvest the Amplified Products

1. When the mRNA sequencing preparation script has finished, tap **EJECT** to remove the IFC from the instrument (see Note 17).
2. Transfer the C1 IFC to a post-PCR lab environment.
3. Label a new 96-well plate Diluted Harvest Plate.
4. Aliquot 10 μL of C1 DNA Dilution Reagent into each well of the diluted harvest plate.

Table 1
Thermal cycling protocols

	Temperature (°C)	Time	Cycles
Lysis program	72	3 min	
Reverse transcription program	4	10 min	
	25	1 min	
	42	90 min	
PCR amplification program	70	10 min	
	95	1 min	
	95	20 s	5
	58	4 min	
	68	6 min	
	95	20 s	9
	64	30 s	
	68	6 min	
	95	30 s	7
	64	30 s	
	68	7 min	
	72	10 min	

5. Carefully pull back the tape covering the harvesting inlets of the IFC using the plastic removal tool.
6. Using an 8-channel pipette, pipet the harvested amplicons from the inlets according to Fig. 6, and place them in the diluted harvest plate (*see* **Note 18**). The harvest amplicon dilution will consist of 10 μ L C1 DNA Dilution Reagent and 3 μ L C1 harvest amplicons. Detailed instructions on pipetting the harvested aliquots to the diluted harvest plate:
 - (a) Pipet the entire volume of C1 harvest amplicons out of the left-side wells of the C1 IFC into the 10 μ L of C1 DNA Dilution Reagent in each well of the diluted harvest plate (Fig. 7).
 - (b) Pipet the entire volume of C1 harvest amplicons out of the right-side wells of the C1 IFC into the 10 μ L of C1 DNA Dilution Reagent in each well of the diluted harvest plate (Fig. 8).
 - (c) Pipet the entire volume of C1 harvest amplicons out of the left-side wells of the C1 IFC into the 10 μ L of C1 DNA Dilution Reagent in each well of the diluted harvest plate (Fig. 9).
 - (d) Pipet the entire volume of C1 harvest amplicons out of the right-side wells of the C1 IFC into the 10 μ L of C1 DNA Dilution Reagent in each well of the diluted harvest plate (Fig. 10).

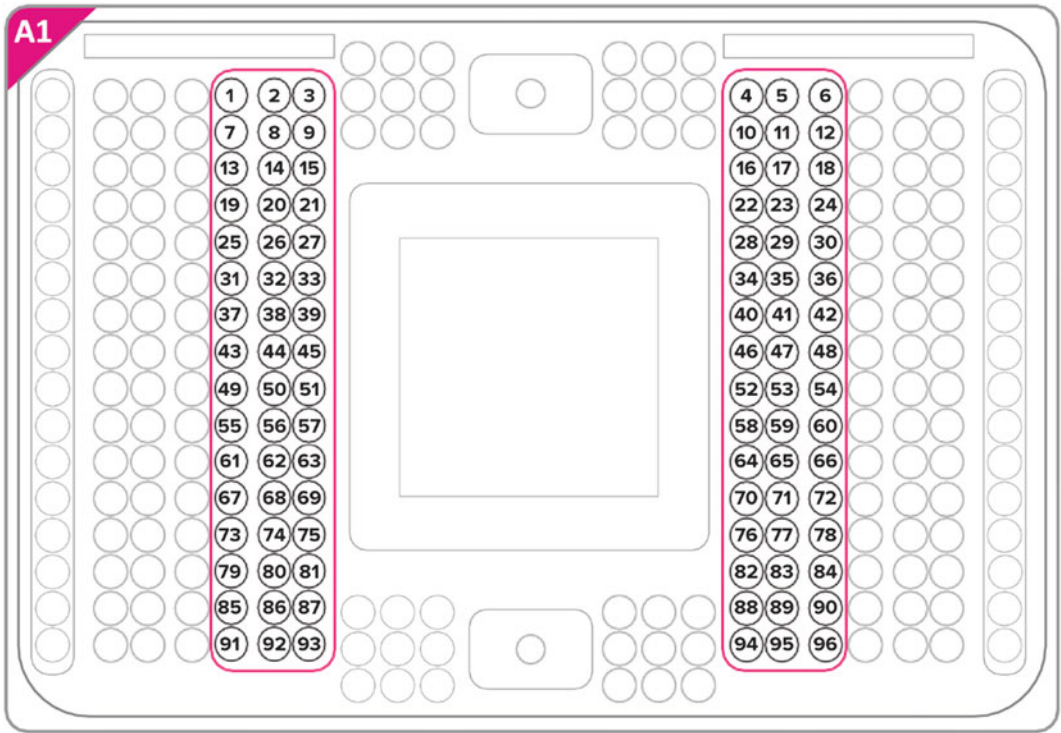


Fig. 6 Pipette map of reaction products on the C1 IFC. Schematically shown is the C1 IFC. Inlets that contain harvest product are numbered

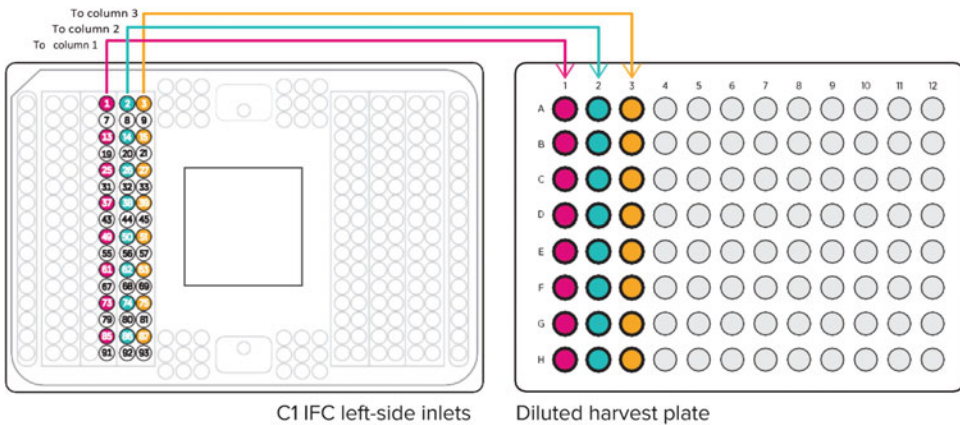


Fig. 7 First three harvest product pipette steps. Schematically shown is the pipetting scheme for harvesting the PCR products from the C1 IFC (*left*) to the 96-well PCR plate (*right*)

- (e) Seal, vortex the harvest plate for 10 s, and then centrifuge it to collect harvest products. After harvesting, material from the capture sites is arranged on the harvest plate (Fig. 11).

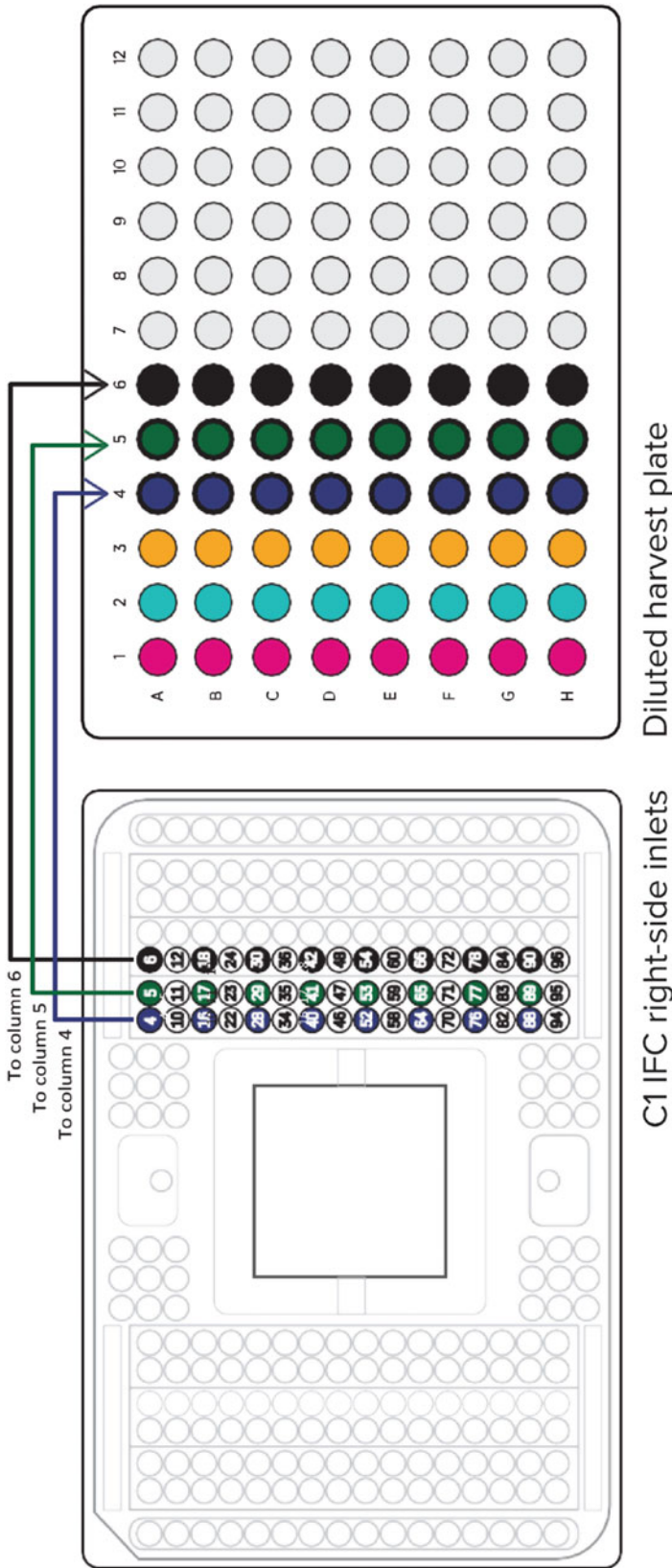


Fig. 8 Fourth, fifth, and sixth pipetting steps. Schematically shown is the pipetting scheme for harvesting the PCR products from the C1 IFC (*left*) to the 96-well PCR plate (*right*)

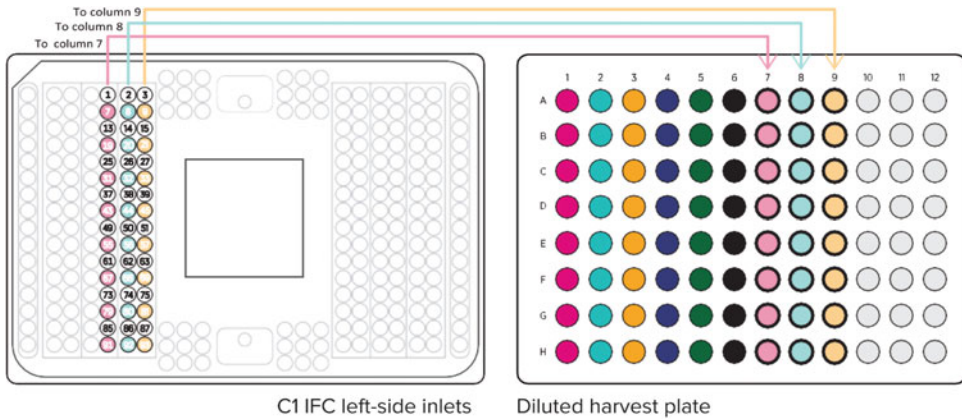


Fig. 9 Seventh, eighth, and ninth pipetting steps. Schematically shown is the pipetting scheme for harvesting the PCR products from the C1 IFC (*left*) to the 96-well PCR plate (*right*)

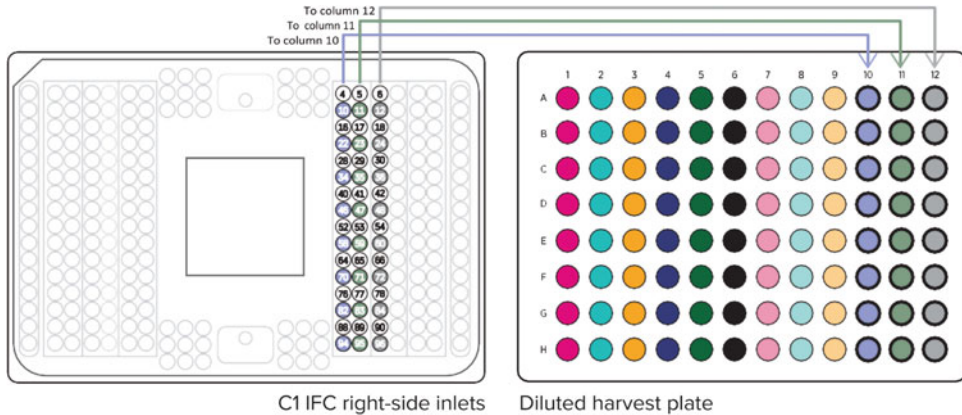


Fig. 10 Tenth, eleventh, and twelfth pipetting steps. Schematically shown is the pipetting scheme for harvesting the PCR products from the C1 IFC (*left*) to the 96-well PCR plate (*right*)

3.9 Quantify and Dilute Harvested cDNA (See Note 19)

1. Label a new 96-well PCR plate “Diluted Samples”.
2. Pipet the appropriate amount of C1 Harvest Reagent to each well of the diluted samples plate as follows per determined sample dilution: 2 μ L (1:2), 4 μ L (1:3), 6 μ L (1:4), 8 μ L (1:5), 10 μ L (1:6), 14 μ L (1:8), 18 μ L (1:10), 22 μ L (1:12).
3. Transfer 2 μ L of the harvest sample from the harvest sample plate to the diluted samples plate.
4. Seal the plate with adhesive film.
5. Vortex at medium speed for 20 s and centrifuge at $367 \times g$ for 1 min.

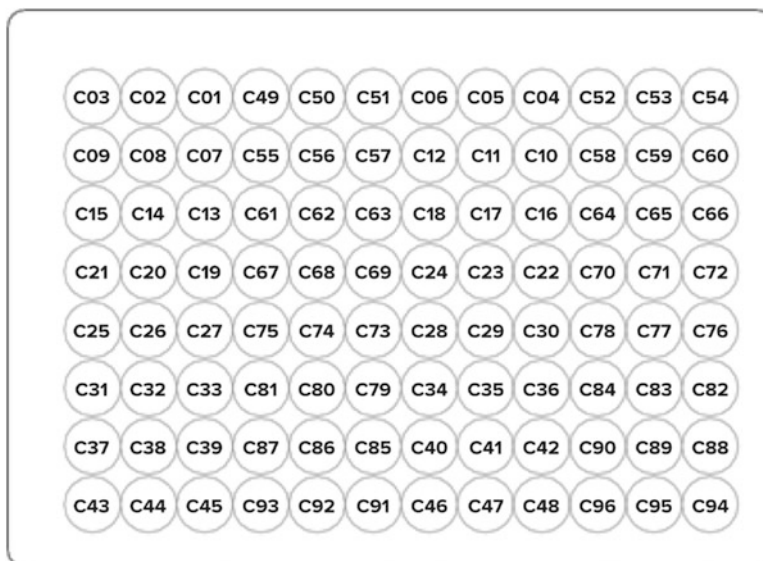


Fig. 11 Sample attribution in 96-well format. Schematically shown is the harvest plate with the attributed samples as they correspond to the C1 IFC (e.g., C03 = capture site 3 is well A1 in harvest plate)

3.10 Prepare cDNA for Tagmentation (See Note 20)

1. After thawing, gently invert the tubes 3–5 times to mix reagents, and then centrifuge tubes briefly to collect the contents.
2. Label a new 96-well PCR plate “Library Prep”.
3. In a 1.5 mL PCR tube, combine the components of the premix: 300 μ L of Tagment DNA Buffer and 150 μ L of Amplicon Tagment Mix (total volumes include 25% overage for 96 samples).
4. Vortex at low speed for 20 s and centrifuge the tube to collect contents.
5. Aliquot equal amounts of premix into each tube of an 8-tube strip.
6. Pipet 3.75 μ L of the premix to each well of the library prep plate using an 8-channel pipette.
7. Pipet 1.25 μ L of the diluted sample from the diluted sample plate to the library prep plate.
8. Seal plate and vortex it at medium speed for 20 s. Centrifuge at 4000 rpm for 5 min to remove bubbles.
9. Place the library prep plate in a thermal cycler and run the following program: 55 $^{\circ}$ C for 10 min and hold at 10 $^{\circ}$ C.
10. Aliquot equal amounts of NT Buffer into each tube of an 8-tube strip. You need 1.25 μ L of NT Buffer for each sample plus 25% overage (150 μ L total for 96 samples).

11. When the sample reaches 10 °C, pipet 1.25 µL of the NT Buffer to each of the tagmented samples to neutralize the samples.
12. Seal plate and vortex at medium speed. Centrifuge at 2200 × *g* for 5 min.

3.11 Amplify the DNA (See Note 21)

1. Aliquot equal volumes of Nextera PCR Master Mix (NPM) into each tube of an 8-tube strip.
2. Pipet 3.75 µL of the aliquoted NPM to each well of the library prep plate using an 8-channel pipette.
3. Select appropriate Index 1 (N7 $_{xx}$) and Index 2 (S5 $_{xx}$) primers for the number of samples in your experiment. Each Index 1 Primer corresponds to a column of the 96-well plate and each Index 2 Primer corresponds to a row.
4. Pipet 1.25 µL of Index 1 Primers (N7 $_{xx}$) to the corresponding well of **each row** of the library prep plate using a 12- or 8-channel pipette.
5. Pipet 1.25 µL of Index 2 Primers (S5 $_{xx}$) to the corresponding well of **each column** of the library prep plate using an 8-channel pipette.
6. Seal the plate with adhesive film and vortex at medium speed for 20 s. Centrifuge at 2200 × *g* for 2 min.
7. Place the plate into a thermal cycler and perform PCR amplification corresponding to the following thermal cycling protocol: 72 °C for 3 min, 95 °C for 30 s, 12 cycles of 95 °C for 10 s, 55 °C for 30 s, 72 °C for 60 s, 1 cycle of 72 °C for 5 min, and then hold at 10 °C (*see Note 22*).
8. Amplified products can be stored at –20 °C for long-term storage.

3.12 Pool and Clean Up the Library

1. Determine the number of samples to be pooled based on desired sequencing depth and sequencer throughput (*see Note 23*).
2. Warm Agencourt AMPure XP beads to room temperature and vortex for 1 min.
3. Make library pools by pipetting the appropriate volume from each sample listed in Table 2 according to the determined number of samples to be pooled.
4. To the pooled library, add the required amount of AMPure XP beads listed in Table 2.
5. Mix well by pipetting up and down five times.
6. Incubate the bead mix at room temperature for 5 min.
7. Place the tube on a magnetic stand for 2 min.

Table 2
Sample volume to be pooled for different pool sizes and AMPure beads required

Number of samples to be pooled	Volume per sample (μL)	Total pool volume (μL)	AMPure bead volume for cleanup (μL) (90% of total pool volume)
8	4	32	29
12	4	48	44
16	2	32	29
24	2	48	44
32	1	32	29
48	1	48	44
96	1	96	87

8. Carefully remove the supernatant without disturbing the beads.
9. Add 180 μL of freshly prepared 70% ethanol and incubate for 30 s on the magnetic stand.
10. Remove the ethanol.
11. Repeat **steps 9 and 10**.
12. Allow the beads to air-dry on bench for 10–15 min.
13. Elute the samples by adding the required volume of C1 DNA Dilution Reagent per number of samples pooled according to Table 3.
14. Vortex the tube for 3 s and incubate it for 2 min at room temperature.
15. Plate the tube on a magnetic stand for 2 min.
16. Transfer the entire volume of supernatant to another PCR tube.

3.13 Repeat Cleanup

1. Add the required amount of AMPure XP beads according to Table 4.
2. Mix well by pipetting up and down five times.
3. Incubate the bead mix 5 min at room temperature.
4. Place the tube on a magnetic stand for 2 min.
5. Carefully remove the supernatant without disturbing the beads.
6. Add 180 μL of freshly prepared 70% ethanol and incubate for 30 s on the magnetic stand.
7. Remove the ethanol.

Table 3
Elution buffer required for libraries pooled from different number of samples

Number of libraries pooled	Volume of C1 DNA Dilution Reagent (volume of original sample pool; μL)
8	32
12	48
16	32
24	48
32	32
48	48
96	96

Elution buffer volume is equal to pooled library volume

Table 4
Elution buffer required for libraries pooled from different number of samples

Number of libraries pooled	AMPure bead volume for cleanup (90% of total pool volume; μL)
8	29
12	44
16	29
24	44
32	29
48	44
96	87

8. Repeat **steps 6** and **7** (*see* **Note 24**).
9. Allow beads to air-dry on bench for 10–15 min.
10. Elute the samples by adding the required volume of C1 DNA Dilution Reagent per number of samples pooled according to **Table 5**.
11. Remove the tube from the magnetic stand and vortex the tube for 3 s.
12. Incubate at room temperature for 2 min.
13. Place the tube on the magnetic stand for 2 min.

Table 5
Final elution buffer required for libraries pooled from different number of samples

Number of libraries pooled	Volume of C1 DNA Dilution Reagent (1.5 × original pool volume; μL)
8	48
12	66
16	48
24	66
32	48
48	66
96	144

14. Carefully transfer the supernatant to another PCR tube labeled “SC Lib”.
15. Perform Agilent Bioanalyzer analysis in triplicate using the Agilent High Sensitivity DNA chip for library size distribution and quantitation. Refer to the Agilent Bioanalyzer user guide for this step.
16. Refer to the Illumina sequencing manual to determine the appropriate library concentration for sequencing.

4 Notes

1. This reagent mix is sufficient for 125 C1 IFCs. Due to the low volume pipetted, we recommend making the mix in bulk and aliquoting for future use. While ArrayControl RNA Spikes contain eight RNA transcripts, we use only three. Alternatively, ArrayControl Spikes can be replaced with External RNA Control Consortium (ERCC) Spike-in RNAs, which are comprised of 92 synthetic polyadenylated RNAs with varying lengths and molecule numbers [6]. Applicable concentrations are cell-type dependent and need to be determined empirically (they can range from 1:20 K to 1:1000 K). Expression data derived from ERCC spike-ins can be used for a number of critical analysis steps, including (but not limited to) data normalization, low-quality data removal, technical noise assessment/removal, and batch-effect correction.
2. Diluted RNA does not store well. Do not dilute RNA for longer than 1 h before loading the IFC. Store only concentrated aliquots long term.

3. You can combine 1 μL of the RNA Spikes mix with 9 μL of Loading Reagent, and then combine 1 μL of the resulting mix with 9 μL of Loading Reagent. Vortex the diluted RNA Spikes mix for 3 s after each dilution.
4. If you are not using RNA spikes, add 1 μL of Loading Reagent instead of the diluted RNA Spikes mix.
5. The Clontech SMARTer kit contains two PCR buffers. Do not use the short amplicon (SA) buffer.
6. After priming the IFC, you have up to 1 h to load the IFC with the C1 system.
7. Keep the dye tubes closed and in the dark as much as possible, because they can hydrolyze over time. When not in use, store in a dark, airtight bag with desiccant pack at -20°C . Cell staining solution may be prepared up to 2 h before loading into the C1 IFC. Keep on ice and protected from light before pipetting into the IFC. Staining small cells (5–10 μm) takes 30 min, and staining medium (10–17 μm) or large (17–25 μm) cells takes 60 min.
8. The quality of the cell suspension is of critical importance to ensure a high number of cells captured on the IFC and a successful workflow. In general, capture efficiency (i.e., the percentage of cells captured on the IFC) depends on cell concentration, cell size, cell buoyancy, fraction of viable/dead cells, and presence of debris. Determining all parameters *prior to* running an experiment on the C1 IFC is recommended. **Concentration:** Cells may be counted by any preferred method. In the absence of an established cell counting protocol, we suggest using the disposable hemocytometer C-Chip by INCYTO. The ratio of cells to suspension reagent may need to be optimized to maximize cell capture, as discussed below. Do not exceed a total of 1000 cells when loading the C1 IFC. Suitable concentrations range from 66 cells/ μL (large cells, 17–25 μm) to 400 cells/ μL (small cells, 5–10 μm). Concentrations refer to cell suspension prior to adding Suspension Reagent. When working with a rare, low-number target cell population, cells can be collected (e.g., by fluorescence-activated cell sorting) directly on the IFC. Please contact the Technical Support Team at Fluidigm for additional information. **Size:** Average size and size distribution of target cells can be determined manually using a hemocytometer or automatically using automated cell-counting instruments (e.g., CountessTM, Vi-CELL[®]). Results of this experiment determine which IFC format to use (5–10 μm , 10–17 μm , or 17–25 μm). Target populations with multimodal size distributions spanning a broad range of sizes may be size-partitioned (e.g., by fluorescence-assisted cell sorting, or FACS) prior to the workflow and run on different-

size IFCs. **Viability:** Including a viability stain, such as trypan blue, will allow for assessment of percentage of viable cells. **Buoyancy:** Routinely the cells are mixed with a cell-type-dependent volume of Cell Suspension Reagent (CSR) to ensure optimal buoyancy of the cells throughout the cell-loading step. Optimal buoyancy may be qualitatively determined with a serial titration of varying concentrations of cells, ranging from 50 to 100%). In a 384-well flat-bottom plate, a total of 10 μL of cells + Cell Suspension Reagent may be plated (start with 5 μL cells +5 μL CSR = 50%, end with 10 μL cells +0 μL CSR = 100%). After minimum waiting (that is, the maximum time it takes for the cell-loading script to finish), microscopically assess cell distribution. Evenly dispersed cells throughout the column of the well (a few cells visible in each focal plane) suggest optimal buoyancy. Cells visible primarily only in one or a few focal planes suggest suboptimal buoyancy (Fig. 2).

9. Vortex the Suspension Reagent for 5 s before use. If Suspension Reagent contains particulates, ensure that they are properly removed by vortexing. Do not vortex cells.
10. The Load and Staining script for small cells (5–10 μm) takes 30 min, and for medium (10–17 μm) or large (17–25 μm) cells, 60 min.
11. For cell types that do not exhibit inhibition, sample preparation may be performed according to Protocol B.
12. Review the Qiagen RNeasy Plus Micro Kit Protocol for proper use and handling of material before proceeding. Some components contain guanidine thiocyanate, which can form highly reactive compounds when combined with bleach. Take special care in handling and disposal.
13. Even though cells are lysed, continue with the lysis reaction as written, because the 3' SMART primer is added to the reaction during this step.
14. This is a potential stopping point. PCR mix and RT reaction products can be stored at 4 $^{\circ}\text{C}$ in a thermal cycler overnight and prepared the following morning.
15. Imaging the IFC presents a critical step of the workflow because it enables visual confirmation of a single, viable cell prior to processing the samples. Low-quality samples, particularly doublets/multiplets (resulting from insufficient dissociation or incorrect capture) can be detected and may be removed at later stages of the workflow (e.g., prior to library preparation or after sequencing). These data points are challenging, if not impossible, to identify based solely on expression data.

16. The mRNA Seq: RT & Amp (1771×/1772×/1773×) script may be run overnight. Approximate run times:
 - **Small-cell IFC:** ~7.75 h (6.5 h for lysis, reverse transcription, and amplification and 1.25 h for harvest).
 - **Medium- and large-cell IFCs:** ~8.5 h (6.5 h for lysis, reverse transcription, and amplification and 2 h for harvest).
17. The IFC may remain in the C1 system for up to 1 h after harvest before you remove products from their inlets.
18. Harvest volumes may vary. Set a pipette to 3.5 μL to ensure entire volume is extracted.
19. cDNA concentrations yielded from the C1 system may vary with cell types and cell treatments. Both the library yield and size distribution also vary with input cDNA/DNA concentrations. To minimize library prep variation and to achieve high library quality, the harvest concentration and dilution must be carefully determined. We suggest using the PicoGreen assay to determine the concentration of cDNA samples; however, other methods can be used. We suggest using the Microsoft Excel[®] worksheet, Single-Cell mRNA Seq PicoGreen Template (Fluidigm), to quantify the library. The optimal concentration for Nextera XT library preparation is 0.1–0.3 ng/ μL . Dilute each sample with the appropriate dilution factor to fall within this range. This can be done with single or multiple dilution steps. If a 384-well fluorometer is not available, an Agilent Bioanalyzer can be used. Samples from a C1 IFC should be run on the Agilent Bioanalyzer with the Agilent High Sensitivity DNA chip. The concentration of each sample is estimated with a size range of 100–10,000 bp. Using the Single-Cell mRNA Seq PicoGreen Template with a Qubit[®] fluorometer is also an option. Input values into Concentration Estimate Table on the Example Results tab of the template.
20. Warm Tagment DNA Buffer and NT Buffer to room temperature. Visually inspect NT Buffer to ensure that there is no precipitate. If there is precipitate, vortex until all particulates are resuspended.
21. Carefully read the Illumina Nextera XT DNA Library Preparation Guide for Index primer selection criteria before proceeding to PCR amplification of the tagged cDNA.
22. Ensure that the thermal cycler lid is heated during the incubation.
23. If preferred, samples can be cleaned up individually prior to pooling.
24. Some beads may be lost during ethanol cleanup.

References

1. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 6(5):377–382. <https://doi.org/10.1038/nmeth.1315>
2. Kolodziejczyk AA, Kim JK, Tsang JC, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, Buhler M, Liu P, Marioni JC, Teichmann SA (2015) Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* 17(4):471–485. <https://doi.org/10.1016/j.stem.2015.09.011>
3. Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, Marques S, Munguba H, He L, Betsholtz C, Rolny C, Castelo-Branco G, Hjerling-Leffler J, Linnarsson S (2015) Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347(6226):1138–1142. <https://doi.org/10.1126/science.aaa1934>
4. Stubbington MJ, Lonnerberg T, Proserpio V, Clare S, Speak AO, Dougan G, Teichmann SA (2016) T cell fate and clonality inference from single-cell transcriptomes. *Nat Methods* 13(4):329–332. <https://doi.org/10.1038/nmeth.3800>
5. Grun D, van Oudenaarden A (2015) Design and analysis of single-cell sequencing experiments. *Cell* 163(4):799–810. <https://doi.org/10.1016/j.cell.2015.10.039>
6. Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, Causton H, Conley MP, Elespuru R, Fero M, Foy C, Fuscoe J, Gao X, Gerhold DL, Gilles P, Goodsaid F, Guo X, Hackett J, Hockett RD, Ikonomi P, Irizarry RA, Kawasaki ES, Kaysser-Kranich T, Kerr K, Kiser G, Koch WH, Lee KY, Liu C, Liu ZL, Lucas A, Manohar CF, Miyada G, Modrusan Z, Parkes H, Puri RK, Reid L, Ryder TB, Salit M, Samaha RR, Scherf U, Sendera TJ, Setterquist RA, Shi L, Shippy R, Soriano JV, Wagar EA, Warrington JA, Williams M, Wilmer F, Wilson M, Wolber PK, Wu X, Zadro R, External RNACC (2005) The external RNA controls consortium: a progress report. *Nat Methods* 2(10):731–734. <https://doi.org/10.1038/nmeth1005-731>

MiSeq: A Next Generation Sequencing Platform for Genomic Analysis

Rupesh Kanchi Ravi, Kendra Walton, and Mahdieh Khosroheidari

Abstract

MiSeq, Illumina's integrated next generation sequencing instrument, uses reversible-terminator sequencing-by-synthesis technology to provide end-to-end sequencing solutions. The MiSeq instrument is one of the smallest benchtop sequencers that can perform onboard cluster generation, amplification, genomic DNA sequencing, and data analysis, including base calling, alignment and variant calling, in a single run. It performs both single- and paired-end runs with adjustable read lengths from 1×36 base pairs to 2×300 base pairs. A single run can produce output data of up to 15 Gb in as little as 4 h of runtime and can output up to 25 M single reads and 50 M paired-end reads. Thus, MiSeq provides an ideal platform for rapid turnaround time. MiSeq is also a cost-effective tool for various analyses focused on targeted gene sequencing (amplicon sequencing and target enrichment), metagenomics, and gene expression studies. For these reasons, MiSeq has become one of the most widely used next generation sequencing platforms. Here, we provide a protocol to prepare libraries for sequencing using the MiSeq instrument and basic guidelines for analysis of output data from the MiSeq sequencing run.

Key words Next generation sequencing, Illumina, Sequencing-by-synthesis, Cluster generation, Data analysis

1 Introduction

Over the last decade, there have been major advancements in the field of next generation sequencing (NGS) technology leading to several important breakthroughs in human disease research, cancer mutation detection, metagenomics, agriculture, and evolutionary biology [1]. The MiSeq system was first released in 2011 as a compact benchtop sequencer using the sequencing-by-synthesis (SBS) technology of Illumina (San Diego, CA). Benchtop sequencing platforms such as MiSeq can sequence libraries rapidly, produce hundreds of gigabases of data, and perform data analysis in a single integrated sequencing run. The MiSeq system includes onboard cluster generation and data analysis and access to BaseSpace[®], the Illumina genomic analysis platform that provides onsite or internet

(cloud)-based, real-time data uploading, data analysis tools, and run monitoring [2]. This chapter describes the MiSeq workflow for library denaturation and dilution, instrument setup, data analysis, and troubleshooting of problems that might occur during the sequencing run.

2 Materials

2.1 Buffers, Solutions, and Reagents

- Illumina HT1 (Hybridization buffer) provided with the MiSeq sequencing kit.
- 0.2 N NaOH (Prepared fresh from 1 N NaOH stock).
- 10 mM Tris–HCl pH 8.5 in 0.1% Tween 20.
- PhiX control.
- Vortexer.
- Microcentrifuge.

2.2 Sequencing Run Kits

- MiSeq reagent kit (V2 or V3) containing MiSeq flow cell, PR2 bottle and reagent cartridge.
- Ethanol and distilled water to wash flow cell.
- Pipettes.
- Centrifuge.

2.3 Equipment

- Hybex microsample incubator.
- Heat block for 1.5 mL tubes.

2.4 Instrument and Data Analysis Software

- MiSeq system.
- Preinstalled software: MiSeq control software (MCS), Real-time analysis (RTA) software, and MiSeq Reporter.

3 Methods

3.1 Instrument Setup

1. Either the MiSeq reagent kit V2 or V3 can be used, depending upon the read length and number of reads required (*see Note 1*).
2. The reagent cartridges (V2 or V3) and HT1 buffer must be removed from $-20\text{ }^{\circ}\text{C}$ and warmed to room temperature before loading on the instrument. Ensure the reagent cartridge is completely thawed and remove any air bubbles by mixing and tapping gently (*see Subheading 3.5*).

3. The compatible kit containing the flow cell and PR2 bottle must be removed from 4 °C and warmed to room temperature for 30 min before preparing the flow cell for loading.
4. A sample sheet (*.csv file; comma-separated values file) containing the information on run setup, analysis of sequencing run, list of library samples pooled, and appropriate corresponding index sequences, is required to set up the run. The file must be copied to the manifest folder on the MiSeq.

3.2 Denaturation and Dilution of Libraries: Standard Normalization Method

The standard normalization method for denaturation and dilution of libraries must be used when the libraries have been normalized using standard library quantification methods (*see Note 2*). Denature libraries with freshly prepared NaOH, ensuring that the final NaOH concentration is not >1 mM. Denatured libraries are then diluted with HT1 buffer (*see Note 3*) [3].

1. For denaturation, 0.2 N NaOH is prepared by diluting stock 1 N NaOH (200 μ L) with distilled water (800 μ L) (*see Note 4*). The diluted 0.2 N NaOH must be used within 12 h of preparation.
2. Combine 5 μ L of 4 nM library and 5 μ L of 0.2 N NaOH, and mix by vortexing, followed by centrifugation at $280 \times g$ for 1 min (For reagent kit V2, 2 nM library could be used for the denaturation and dilution steps) (*see Note 5*).
3. Incubate mixture for 5 min at room temperature.
4. After incubation, add 990 μ L HT1 buffer to dilute the library to 20 pM. Vortex the mixture, and then centrifuge at $280 \times g$ for 1 min.
5. Depending upon user preferences, load between 6 and 20 pM of diluted libraries (600 μ L) into the reagent cartridge for sequencing (Table 1; *see Note 6*).

Table 1
Preparing different concentrations of sequencing library pools

Pool loading concentration	20 pM Denatured library (μ L)	HT1 buffer (μ L)
6 pM	180	420
8 pM	240	360
10 pM	300	300
12 pM	360	240
15 pM	450	150
20 pM	600	0

3.3 PhiX Control, Denaturation, and Dilution for Low Diversity Samples

1. The concentration of the PhiX control supplied by illumina is 10 nM. For denaturation of PhiX, mix 2 μL of the 10 nM stock with 3 μL of 10 mM Tris-Cl, pH 8.5 in 0.1% Tween 20. The resulting concentration of PhiX is 4 nM and is stable for up to 12 h.
2. For denaturation of the working solution of the PhiX control, mix 5 μL of the 4 nM PhiX control with 5 μL of freshly prepared 0.2 N NaOH, vortex briefly, and then centrifuge at $280 \times g$ for 1 min.
3. Incubate mixture at room temperature for 5 min to denature PhiX.
4. Dilute the denatured 4 nM PhiX control library (10 μL) with 990 μL of pre-chilled HT1 and mix by inverting. The resulting concentration is 20 pM, which is used for MiSeq Reagent kit v3 without further dilution. The 20 pM PhiX control can be stored at $-15\text{ }^{\circ}\text{C}$ to $-25\text{ }^{\circ}\text{C}$ up to 3 weeks.
5. When using MiSeq reagent kit v2, dilute the denatured 20 pM PhiX control to 12.5 pM by mixing 375 μL 20 pM PhiX with 225 μL prechilled HT1. The resulting mix produces optimal cluster density when using v2 reagents.
6. For low diversity libraries, at least 5% PhiX control is spiked-in by mixing denatured and diluted libraries of 570–30 μL of denatured and diluted PhiX control (*see Note 7*).
7. For most libraries, 1% PhiX control spike-in is used by mixing 6 μL of denatured and diluted PhiX control with 594 μL of denatured and diluted libraries.

3.4 Denaturation and Dilution of Libraries: Bead-Based Normalization Method

The bead-based normalization method for denaturation and dilution of libraries is dependent on library preparation types that include normalization using bead-based procedures. Examples of this type of normalization method are Nextera XT library pools and certain Amplicon library pools.

1. Preheat an incubator to $98\text{ }^{\circ}\text{C}$.
2. Combine 6–10 μL of amplicon library pools with 594–590 μL of pre-chilled HT1 buffer, respectively, for a total volume of 600 μL . For library pools from Nextera XT, mix 24 μL of library pool with 576 μL of HT1 buffer, again the final volume is 600 μL (Table 2).
3. Briefly vortex mixture, and then centrifuge at $280 \times g$ for 1 min.
4. Incubate diluted library pools at $98\text{ }^{\circ}\text{C}$ for 2 min, and then cool immediately on ice for 5 min.

Table 2
Sample library preparation for bead-based normalization method

Library normalization method	Library pool (μL)	Prechilled HT1 buffer (μL)
Amplicon	6	594
	7	593
	8	592
	9	591
	10	590
Nextera XT	24	576

- The denatured and diluted libraries are now ready to be loaded on to the MiSeq cartridge, unless the library pools have low diversity. For low diversity libraries, similar steps as detailed above should be followed, using the PhiX control before loading on to the sequencer (*see* Subheading 3.3).

3.5 Sequencing Setup

The denatured and diluted libraries, prepared by either the standard or bead-based normalization method, are ready for loading onto the reagent cartridge and the sequencing run can be initiated.

- Invert the thawed reagent cartridge (either V2 or V3 kit) ten times to mix contents in the reservoir positions of the cartridge and inspect cartridge for air bubbles or precipitates in positions 1, 2, and 4.
- Tap cartridge gently to dispel air bubbles.
- Load the denatured and diluted library pool of 600 μL (with or without PhiX control) into the reservoir position labeled “Load samples” by piercing the foil seal with tip.
- The prepared cartridge is now ready for sequencing.
- Before loading the MiSeq flow cell on to the sequencer, it must first be rinsed using distilled water and dried with lint-free lens cleaning tissue, taking care not to disturb the black flow cell port gasket.

3.6 MiSeq Control Software and Starting the Run

The MiSeq control software (MCS v2.3) provides step-by-step instructions for loading the flow cell, PR2 bottle, waste bottle, reagent cartridge with libraries and input of manifest file location containing the sample sheet information. The MCS reviews the run parameters by performing pre-run check of all run components, disk space, and network connections. After the completion of the pre-run check, the sequencer is ready to be started. The run time of the sequencer depends on the selection, whether sequencing is single- or paired-end, and read length, and can range from 4 to 56 h.

4 Diagnosis and Troubleshooting of MiSeq Setup and Sequencing

Illumina's library and sequence-by-synthesis technology is based on three steps: preparation of libraries, generation/amplification of clonal clusters from libraries, and deep, massively parallel sequencing. The sequencing performance on the MiSeq system, in terms of data quality and total data output yield, depends on various factors such as preparation of libraries, quantification of the libraries, denaturation and dilution of libraries, density of clonal clusters, loading concentration, and sequencing run parameters [4].

One of the important parameters affecting run quality, passing filter reads, Q30 scores, and total output data yield is cluster density. Underclustering and overclustering can lead to poor performance of the sequencing run, lower Q30 scores, introduction of sequencing artifacts, increases in sequencing errors, and negative effects to the total data output yield. The goal of any sequencing run is to have good cluster densities without overclustering and maximum total data output. Below are some of the common problems that occur as a result of underclustering or overclustering and approaches to diagnose them. Both troubleshooting and diagnosis examples are summarized in Table 3.

4.1 Low Q30 Scores

Q30 scores are affected by overloaded signal intensities, which produces decreased base intensity to background ratios and leads to errors in base calling and overall lower data quality. The intensity analysis tab and % Q30 plot on the sequencing analysis viewer (SAV) provides information on the influence of cluster density on Q30 scores to diagnose the error. The lower quality may indicate poor template generation and may affect either Read 1 or Read 2.

4.2 Low Passing Filter Clusters and Data Output Yield

Cluster passing filter percentage (% PF) indicates purity of the signals in each cluster. The decrease in % PF may be due to poor template generation. The density box plot on the SAV compares raw cluster density to % PF. The closer the plot between raw cluster density to % PF, the better the sequencing run. When the % PF decreases, the plot between raw cluster density to % PF will appear farther apart. Low % PF results in reduced output data (*see Note 8*).

4.3 Library Quality

Library preparation is the most important step for generating high quality libraries. High quality libraries produce good template generation and accurate demultiplexing. Errors that can occur during library preparation, such as cleanup steps, can lead to the presence of adapter dimers, primer dimers, partial library constructs, and incorrect indexing of libraries, which affect library quantification and subsequent cluster efficiency. Errors in library cleanup steps can be identified using profiles obtained with the Bioanalyzer or Tapesation methods.

Table 3
Diagnosis and correction of common problems occurring during MiSeq runs

Analysis of problem		Diagnosis using SAV	Correction
Low Q30 scores	Affects signal intensities leading to error in base calling, lower data quality	Intensity and Q30 tab profile on SAV indicates poor template generation	Adjust loading concentration between 6 and 20 pM
Low passing filter (% PF)	Overlapping of clusters indicating poor signal purity	In the density box plots, % PF and raw cluster density appear further apart	Adjust loading concentration between 6 and 20 pM
Library quality	Poor cleanup of libraries lead to presence of adapter dimers, primer dimers, partial libraries lead to lower efficiency	Tapestation profile verifies the quality and purity of libraries	Appropriate cleanup of the libraries
Library quality	Excess NaOH concentration leads to error in library denaturation	NaOH must be prepared fresh, pH >12.5 Final concentration of NaOH in diluted libraries must be <1 mM	Tris-HCl can be used to neutralize the pH
Library quantification	Inaccurate library quantification is one of the common cause of underclustering and overclustering	qPCR quantification, qubit and tapestation profile	qPCR measures exact functional library fragments
Flow cell loading	Lower or higher loading concentration leads to underclustering or overclustering		Load between 6 and 20 pM

4.4 NaOH

The quality and concentration of the NaOH solution used during the denaturation step also affects library quality, cluster generation, and cluster density. The NaOH solution must be prepared fresh and have pH > 12.5, and the concentration of NaOH in the diluted library samples must be <1 mM. For some libraries, Tris-HCl can be used to neutralize the pH (*see Note 3*).

4.5 Library Quantification

Library quantification has been reported as one of the most common causes of underclustering or overclustering of libraries. The selection of an appropriate quantification method is very important. Quantitative PCR (qPCR), picogreen/qubit method, and the Bioanalyzer/Tapestation, and NanoDrop instruments are effective methods for library quantification. The most effective method for library quantification is qPCR, which measures only functional library fragments, while ignoring primer dimers, unindexed library

Table 4
Summary of the MiSeq sequencing run specifications

	Specifications	MiSeq reagent kit v2	MiSeq reagent kit v3
Cluster generation	Read length	1 × 36 bp	2 × 75 bp
		2 × 25 bp	2 × 300 bp
		2 × 150 bp	
		2 × 250 bp	
	Number of cycles	50, 300 And 500 cycles	150 And 600 cycles
	Total run time	4–39 h ^a	21–56 h ^a
	Total output data yield	0.54 Gb–8.5 Gb ^a	3.3–15 Gb ^a
Sequencing run parameters	Passing filter reads	12–30 Million reads	22–50 Million reads
	% Passing filter (% PF)	>90%	>85%
	Quality scores (Q30)	>90% Q30 for 1 × 36 bp and 2 × 25 bp	>85% Q30 for 2 × 75 bp
		>80% Q30 for 2 × 150 bp	>70% Q30 for 2 × 300 bp
		>75% Q30 for 2 × 250 bp	
Raw cluster density (K/mm ²)	865–965	1200–1400	

^aDepends on the read length, the lower the read length, the faster the run time and the lower the output data yield

fragments, and free nucleotides that may be present in the sample. The selection of the library quantification method also depends on the type of library preparation kit, as discussed above (*see Note 9*).

4.6 Flow Cell Loading Concentration

Optimal loading concentration leads to good cluster density and sequencing data. The loading concentration directly affects the raw cluster density and data output. The optimal raw cluster densities are 865–965 K/mm² for v2 MiSeq kits and 1200–1400 K/mm² for v3 MiSeq kits (Tables 2 and 4). An optimal MiSeq run on sequencing analysis viewer (SAV) is shown in Fig. 1 [5].

In summary, the MiSeq system provides a complete end-to-end sequencing solution by integrating cluster generation, amplification, sequencing, and data analysis into a single instrument. The result is low sequencing error, high quality base-by-base accuracy, and elimination of repetitive sequence regions or homopolymers. MiSeq has a fast turnaround time, with less than 4 h of run time and 90 min of library preparation time.

5 Notes

1. V2 kits will produce 12–15 M single reads and 24–30 M paired-end reads. The v2 kit offers three different cycle configurations, the 50-cycle kit, 300-cycle kit and 500-cycle kit with enough chemistry to sequence dual-indexed libraries. The v3 kits produce up to 25 M single reads and offers two different

Flow Cell: 00000000-AJ7JL Extracted: 168 Called: 168 Scored: 168

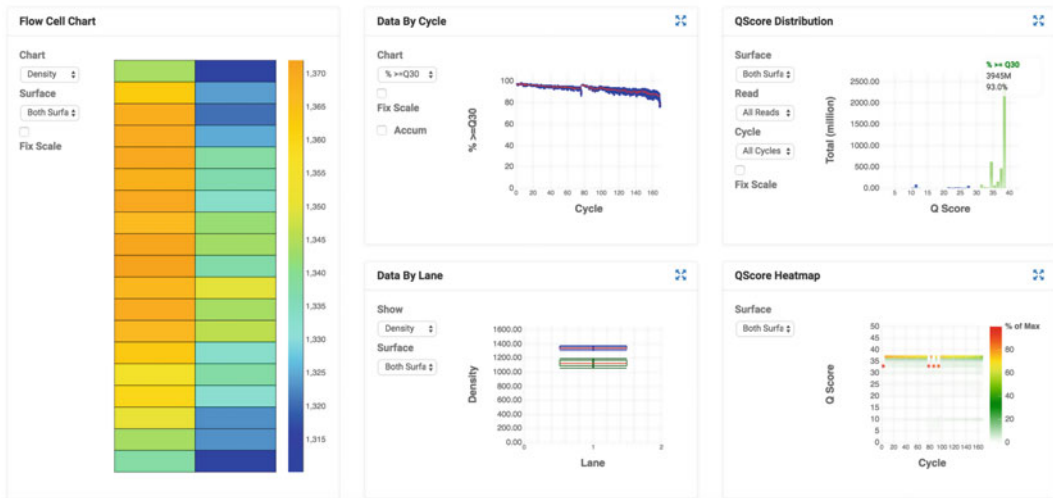


Fig. 1 Example of successful MiSeq sequencing run using a Sequencing Analysis Viewer (SAV) software. SAV viewer provides information on cluster density, Q30 scores, PF. An ideal MiSeq run on V3 kit has 1200–1400 K/mm² cluster density, >75% Q30 and >75% PF

cycle configurations, the 600-cycle kit (2×300 bp) and the 150-cycle kit (2×75 bp), both allowing for dual-indexed libraries.

- Standard library quantification is referencing Bioanalyzer, qubit, qPCR, or a combination of whatever QC metric fits best into the workflow. If libraries are normalized with a bead-based method in the library build protocol, the “Bead-Based Normalization” will need to be used to denature and dilute samples, as those libraries will already be single-stranded.
- It is imperative that the final library/pool dilution has no more than 1 mM NaOH, as higher concentrations of NaOH inhibit cluster formations by blocking hybridization to the flowcell. This will decrease cluster density and result in less data.
- Dilute NaOH in 1 mL volume so that 1 N NaOH can be accurately pipetted. As noted above, inaccurate NaOH concentration inhibits cluster formation.
- If the library pool is low in concentration (less than 2 nM), one of the following options can be employed: (1) the volume of the library pool can be reduced using the SpeedVac concentrator, which will increase the concentration of the pool, (2) use SPRI bead cleanup at a 2:1 ratio with low concentration library pool and eluting at lower volume of solution, (3) the denature and dilute protocol for low concentration library pool can be used by denaturation with NaOH and neutralization using

either HT buffer or 200 mM Tris–HCl pH 7, making sure the final concentration is less than 1 mM.

6. Ideally, 8 pM loading concentration is recommended to achieve higher Q30 scores, passing filter reads, and good quality reads, but optimization for individual MiSeqs and different library preparation protocols will need to be considered.
7. For low diversity libraries where a significant number of reads have the same sequence and therefore lack variation, base compositions of the reads are no longer random. To increase the diversity (>25%), PhiX control must be spiked-in to the denatured and diluted libraries before loading.
8. When looking at SAV, a gradient of the density of the flowcell from top to bottom should be observed. The density should be higher at the ports (bottom of picture of flowcell) and lower toward the end of the flowcell (top of the picture of the flowcell). If the density gradient is flipped, there are a higher number of clusters at the back of the flowcell compared to the front of the flowcell. This is a good indication of overclustering (Fig. 1) [5].
9. The most important part of library quantification is consistency. Labs choose different quantification methods that suit individual workflow and time constraints, and finding what works best for an individual lab is critical. Be willing to optimize library protocols on each Illumina machine with whichever quantification method works best for your lab.

References

1. Bomar L, Maltz M, Colston S, Graf J (2011) Directed culturing of microorganisms using metatranscriptomics. *MBio* 2:e00012–11
2. Illumina Sequencing Technology (2010) In *Technology Spotlight: Illumina Sequencing*. https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf
3. Illumina MiSeq system: Denature and dilute libraries guide (Jan 2016) https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-denature-dilute-libraries-guide-15039740-01.pdf
4. Illumina MiSeq system guide (Sept 2015) https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-system-guide-15027617-01.pdf
5. Illumina Sequencing analysis viewer software user guide (Oct 2014) https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/sav/sequencing-analysis-viewer-user-guide-15020619-f.pdf

Chapter 13

Methods for CpG Methylation Array Profiling Via Bisulfite Conversion

Fatjon Leti, Lorida Llaci, Ivana Malenica, and Johanna K. DiStefano

Abstract

DNA methylation is a key factor in epigenetic regulation, and contributes to the pathogenesis of many diseases, including various forms of cancers, and epigenetic events such as X inactivation, cellular differentiation and proliferation, and embryonic development. The most conserved epigenetic modification in plants, animals, and fungi is 5-methylcytosine (5mC), which has been well characterized across a diverse range of species. Many technologies have been developed to measure modifications in methylation with respect to biological processes, and the most common method, long considered a gold standard for identifying regions of methylation, is bisulfite conversion. In this technique, DNA is treated with bisulfite, which converts cytosine residues to uracil, but does not affect cytosine residues that have been methylated, such as 5-methylcytosines. Following bisulfite conversion, the only cytosine residues remaining in the DNA, therefore, are those that have been methylated. Subsequent sequencing can then distinguish between unmethylated cytosines, which are displayed as thymines in the resulting amplified sequence of the sense strand, and 5-methylcytosines, which are displayed as cytosines in the resulting amplified sequence of the sense strand, at the single nucleotide level. In this chapter, we describe an array-based protocol for identifying methylated DNA regions. We discuss protocols for DNA quantification, bisulfite conversion, library preparation, and chip assembly, and present an overview of current methods for the analysis of methylation data.

Key words DNA, Methylation, CpG, Epigenetics, Bisulfite conversion

1 Introduction

In 1975, Riggs proposed a model that DNA methylation of cytosine residues in the context of CpG dinucleotides should affect binding of regulatory proteins [1]. Since then, pioneering work has shown that DNA methylation directly silences genes, often-times in heritable patterns and with varying levels of regulation [1, 2]. With the advancement of technology and a better understanding of methylation processes, numerous subsequent studies have implicated methylation in the molecular etiologies of many diseases and genetic events, including neurodegenerative and cardiovascular diseases, diabetes and insulin secretion, various forms of

cancers, X-inactivation, cellular differentiation and proliferation, and embryonic development [3–8].

Epigenetic regulation occurs through a highly conserved enzymatic mechanism that incorporates histone deacetylases, methyl-binding proteins, and DNA methyltransferases (DNMTs) [9]. The 5-methylcytosine (5mC) residue within a CpG sequence context is a highly conserved epigenetic marker in plants, fungi, and animals [10]. Approximately 60–80% of the 28 million CpG sites in the human genome are methylated [11], and given the role of DNA methylation in a diverse array of biological roles and gene regulation, the investigation and identification of CpG sites is important for a better understanding of disease pathogenesis and progression.

The identification and characterization of CpG sites is a growing area of research and there are many platforms available for assessment of DNA methylation, each with advantages and disadvantages in specific applications [12, 13]. Bisulfite sequencing technology is considered the gold standard for detection of methylated DNA due to the qualitative and quantitative ability to identify 5mC at a single base pair resolution. Initially developed by Frommer et al. [13], this method converts unmethylated cytosine residues into uracil residues, which are then displayed as thymines on the resulting amplified sense strand in single-stranded DNA sequencing. In contrast, 5mCs do not respond to bisulfite treatment and will be displayed as cytosine residues on the amplified sense strand, therefore, allowing them to be distinguished from unmethylated cytosine residues [14].

Although new approaches for assessing methylation status are continuously emerging, a popular bisulfite conversion method is methyl-DNA immunoprecipitation (MeDIP), which involves denaturation and precipitation of cleaved DNA using a 5mC antibody, followed by sequencing [15]. Another commonly used technology, shotgun sequencing, has the ability to achieve single-base resolution of bisulfite sequencing, although it cannot distinguish between 5-hydroxymethylcytosine and 5mC [16]. The HumanMethylation450 BeadChip array (Illumina; San Diego, CA) allows the analysis of 19,755 unique CpG islands and 3091 probes at non-CpG sites. The Infinium HumanMethylation450 BeadChip array (Illumina) can interrogate the methylation status of more than 450,000 CpGs throughout the genome, 19,755 unique CpG islands with additional coverage in shore regions and miRNA promoters, as well as 3091 probes at non-CpG sites [17]. Although not as comprehensive as sequencing-based approaches, the 450k array provides simpler analysis and interpretation. Some currently popular methods for 450k array data analysis include *methyllumi*, *minfi*, *wateRmelon*, *RnBeads*, *ChAMP*, and *COHCAP*, which are discussed in greater detail below.

The purpose of this chapter is to provide a detail protocol for performing array-based interrogation of CpG methylation using human genomic DNA. We start with a description of DNA

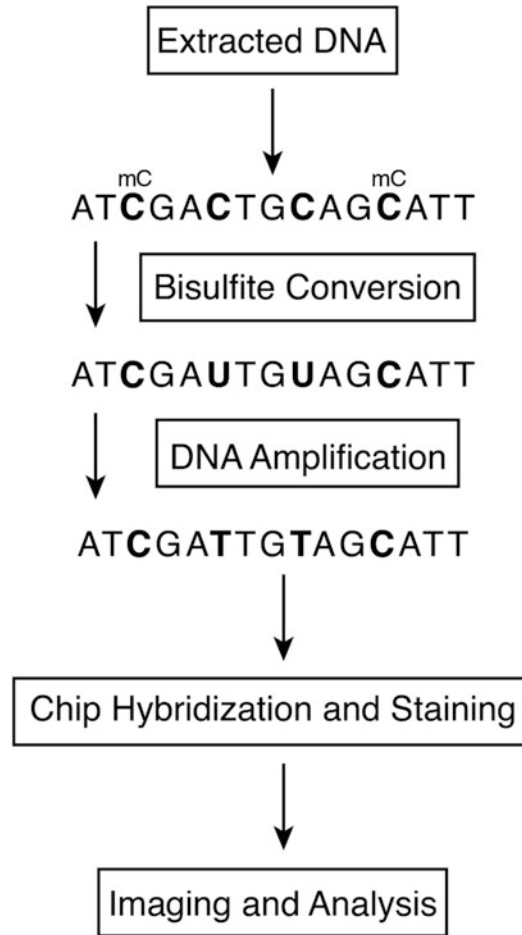


Fig. 1 Schematic representation of workflow to identify gene methylation in conjunction with Illumina Infinium HumanMethylation450 BeadChip Kit assay

quantification using Quant-iT PicoGreen, and follow with a protocol for bisulfite conversion that includes techniques for library preparation, chip assembly, and chip analysis on the Illumina platform (Fig. 1). We also present an approach for sample analysis and provide an overview of current packages for the analysis of methylation data.

2 Materials

2.1 PicoGreen Assay

1. DNA extracted from samples of interest.
2. Quant-iT™ PicoGreen dsDNA Reagent Kit (Thermo Fisher Scientific; Waltham, MA).
3. Black plate with clear flat bottom.
4. 1× Tris-EDTA (TE) buffer.
5. Spectrofluorometer.

2.2 Bisulfite Conversion

1. 500 ng of extracted DNA.
2. EZ DNA Methylation™ Kit (Zymo Research; Irvine, CA).
3. DNA/RNase/PCR inhibitor-free plates and tubes.
4. Thermocycler.

2.3 Methylation Array

1. Infinium HumanMethylation450 BeadChip Kit (Illumina; San Diego, CA).
2. 96-well 0.2 mL skirted microplate.
3. 960 well 0.8 mL microplate (MIDI).
4. 150-mL reservoirs.
5. Heat sealing foil sheets.
6. 0.1 N NaOH.
7. 100% ethanol.
8. 100% 2-propanol.
9. 95% formamide/1 mM EDTA.
10. Wash rack.
11. Infinium standard glass back plates and spacer (Illumina; San Diego, CA).
12. Illumina Hybridization Oven (Illumina; San Diego, CA).
13. Hyb chamber, gasket, and insert (Illumina; San Diego, CA).
14. Tecan.
15. iScan or HiScan (Illumina; San Diego, CA).

3 Methods

3.1 PicoGreen Assay

1. Prepare a 2 µg/mL working solution of the lambda DNA standard, by diluting the 100 µg/mL sample provided in the Quant-iT™ PicoGreen dsDNA Reagent Kit in TE buffer. For a five-point standard curve ranging from 1 ng/mL to 1 µg/mL, proceed to **step 3**. For a low-range standard curve ranging from 25 pg/mL to 25 ng/mL, prepare a 40-fold dilution of the 2 µg/mL DNA solution to yield a 50 ng/mL DNA stock solution and proceed to **step 6**.
2. Prepare an aqueous working solution of the Quant-iT PicoGreen reagent by making a 200-fold dilution of the concentrated DMSO solution in TE (*see Note 1*).
3. For the high-range standard curve, dilute the 2 µg/mL DNA stock solution with TE buffer as shown in Table 1 and add 1.0 mL of aqueous solution of Quant-iT PicoGreen reagent to each cuvette. Mix well and incubate for 2–5 min at room temperature, protected from light.

Table 1
Protocol for preparing a high-range standard curve

TE (μL)	DNA stock (μL) ^a	Diluted PicoGreen reagent (μL)	Final DNA concentration
0	1000	1000	1 $\mu\text{g}/\text{mL}$
900	100	1000	100 ng/mL
990	10	1000	10 ng/mL
999	1	1000	1 ng/mL
1000	0	1000	Blank

^a2 $\mu\text{g}/\text{mL}$

Table 2
Protocol for preparing a low-range standard curve

TE (μL)	DNA stock (μL) ^a	Diluted PicoGreen reagent (μL)	Final DNA concentration
0	1000	1000	25 ng/mL
900	100	1000	2.5 ng/mL
990	10	1000	250 pg/mL
999	1	1000	25 pg/mL
1000	0	1000	Blank

^a50 ng/mL

4. Measure the sample fluorescence using a spectrofluorometer or fluorescence microplate reader and standard fluorescein wavelengths (excitation ~ 480 nm, emission ~ 520 nm) (*see Note 2*).
5. Subtract the fluorescence value of the reagent blank from readings for all of the samples.
6. For the low-range standard curve, dilute the 50 ng/mL DNA stock solution with TE buffer in disposable cuvettes, as shown in Table 2. Add 1.0 mL of the aqueous working solution of the Quant-iT PicoGreen reagent to each cuvette. Mix well and incubate for 2–5 min at room temperature, protected from light. Continue with **steps 2** and **3**. Adjust the fluorometer gain to accommodate the lower fluorescence signals.

3.1.1 Sample Analysis

1. Dilute the experimental DNA solution in TE to a final volume of 1.0 mL in disposable cuvettes or test tubes (*see Note 3*).
2. Add 1.0 mL of the aqueous working solution of the Quant-iT PicoGreen reagent to each sample. Incubate for 2–5 min at room temperature, protected from light.

3. Measure the fluorescence of the sample using the instrument parameters that correspond to those used when generating your standard curve (*see Note 4*).
4. Subtract the fluorescence value of the reagent blank from those of all the samples. Determine the DNA concentration of the sample from the standard curve generated in DNA standard curve.

**3.2 Bisulfite
Conversion Using
the EZ DNA
Methylation™ Kit**

1. Prepare the CT Conversion Reagent by combining 750 μL water and 210 μL M-Dilution Buffer to a tube of CT Conversion Reagent. Mix at room temperature with frequent vortexing or shaking for 10 min (*see Note 5*).
2. Prepare the M-Wash Buffer by adding 24 mL of 100% ethanol to the 6 mL M-Wash Buffer concentrate (D5001) or 96 mL of 100% ethanol to the 24 mL M-Wash Buffer concentrate (D5002).
3. Add 5 μL of M-Dilution Buffer to the DNA sample and adjust the total volume to 50 μL with water. Mix by flicking or pipetting up and down.
4. Incubate the sample at 37 °C for 15 min.
5. Add 100 μL of the CT Conversion Reagent to each sample and mix.
6. Incubate in the dark at 50 °C for 12–16 h.
7. Incubate at 0–4 °C for 10 min.
8. Add 400 μL of M-Binding Buffer to a Zymo-Spin IC™ column and place the column into the collection tube provided in the kit.
9. Load the Sample from **step 7** into the Zymo-Spin IC Column containing the M-Binding Buffer. Close the cap and mix by inverting the column several times.
10. Centrifuge at full speed ($\geq 10,000 \times g$) for 30 s. Discard flow-through.
11. Add 100 μL of M-Wash Buffer to the column. Centrifuge at full speed for 30 s.
12. Add 200 μL of M-Desulfonation Buffer to the column and let stand at room temperature (20–30 °C) for 15–20 min. Following incubation, centrifuge at full speed for 30 s.
13. Add 200 μL of M-Wash Buffer to the column. Centrifuge at full speed for 30 s. Repeat this step.
14. Place the columns in a 1.5 mL microcentrifuge tube. Add 10 μL of M-Elution Buffer directly to the column matrix. Centrifuge at full speed to elute the DNA.

At this point, the DNA can be analyzed or stored at temperatures ≤ -20 °C. For long-term storage (e.g., over a year), store DNA at temperatures ≤ 70 °C (*see Note 6*).

3.3 Methylation Array

3.3.1 The Following Steps Describe the Transfer of Bisulfite-Converted DNA to a MSA4 Plate, the Denaturation and Neutralization of Samples, and the Preparation of Samples for Amplification

1. Preheat the Illumina Hybridization Oven to 37 °C in a post-PCR area.
2. Thaw the MA1, RPM, and MSM reagents to room temperature.
3. Apply a MSA4 barcode label to a new 96-well 0.8 mL microplate (MIDI) (*see Note 7*).
4. Select “MSA4 Tasks” at the robot PC and select the WG#-BCD plate type (midi or TCY), making sure that the “Use Barcodes” checkbox is cleared, and enter the number of DNA samples (48 or 96) in the “Basic Parameters” field.
5. Remove caps from the MA1, RPM, and MSM tubes, and then place them in the robot standoff tube rack according to the bed map.
6. Vortex the sealed plate at 1600 rpm for 1 min, and centrifuge at $280 \times g$ for 1 min.
7. Add 15 mL 0.1 N NaOH to the quarter reservoir, and place the reservoir on the robot bed according to the bed map.
8. Place the WG#-BCD and MSA4 plates on the robot bed according to the bed map, and click “Run” on the robot PC. Then, select the batch you want to run, and click “OK”. Click “OK” again to confirm the required DNAs.
9. After the robot adds the 0.1 N NaOH to the DNA in the MSA4 plate, follow the instructions at the prompt.
10. Seal the plate with a cap mat, vortex at 1600 rpm for 1 min, and centrifuge at $280 \times g$ at 22 °C for 1 min (*see Note 8*).
11. Place the MSA4 plate back to the robot bed, and click “OK”.
12. When the process is done, seal the MSA4 plate with a cap mat, invert it at least ten times to mix contents, and centrifuge at $280 \times g$ for 1 min.
13. Select “Infinium HD Methylation | Incubate MSA4” in the Illumina LIMS left pane.
14. Scan the barcode of the MSA4 plate, click “Verify”, and then “Save”.
15. Incubate in the Illumina Hybridization Oven for 20–24 h at 37 °C.

3.3.2 *The Following Steps Describe the Enzymatic Fragmentation of the Amplified DNA Samples*

1. Preheat the heat block with the midi plate insert to 37 °C.
2. Thaw FMS tubes to room temperature and gently invert at least ten times to mix contents.
3. If you plan to resuspend the MSA4 plate, remove the RAI reagent from the freezer to thaw.
4. Pulse-centrifuge the MSA4 plate at $280 \times g$, remove the cap, and place it on the robot bed according to the bed map. At the robot PC, select “MSA4 Tasks|Fragment MSA4”. Make sure that the “Use Barcodes” checkbox is cleared, and in the “Basic Parameters” pane, change the value for number of MSA4 plates and number of DNA samples per plate to incubate to correspond with the number of samples (*see Note 9*).
5. Place the MSA4 plate on the robot bed according to the bed map.
6. Place FMS tubes in the according bed map, remove the cap, and click “Run”. Click “OK” when finished, remove the plate, and seal it with a cap mat.
7. Vortex at 1600 rpm for 1 min, pulse-centrifuge at $280 \times g$, and place on the 37 °C heat block for 1 h.

At this point the plate can be left in the 37 °C heat block until ready to proceed, or sealed and stored at –25 °C to –15 °C for more than 24 h.

3.3.3 *PM1 and 2-Propanol Are Added to the MSA4 Plate to Precipitate the DNA Samples*

1. Preheat the block to 37 °C, thaw the MSA4 plate to room temperature, and then pulse-centrifuge at $280 \times g$.
2. Thaw the PM1 reagent to room temperature and gently invert to mix at least ten times.
3. Select “MSA4 Tasks|Precip MSA4” at the robot PC, and make sure the “Use Barcodes” checkbox is cleared.
4. Place the plate in the appropriate robot bed.
5. Place a half reservoir in the reservoir frame, and add 1 tube PM1 for 48 samples, or two tubes for 96 samples. Also, place a full reservoir in the reservoir frame, and add 20 mL 2-propanol for 48 samples, or 40 mL for 96 samples, and then click “Run”. Make sure the “Use Barcodes” checkbox is selected.
6. When prompted, remove the plate from the robot bed, seal it with the same cap mat that was removed earlier, vortex at 1600 rpm for 1 min, incubate at 37 °C for 5 min and pulse-centrifuge at $280 \times g$ (*see Note 10*).
7. Remove and discard the cap mat, place the plate back in the robot bed, and click “OK”.
8. When prompted, seal the plate with a new, dry cap mat. Invert the plate ten times to mix, incubate at 4 °C for 30 min,

centrifuge at $3000 \times g$ at $4\text{ }^{\circ}\text{C}$ for 20 min, and immediately remove the MSA4 plate from the centrifuge. Remove and discard the cap mat.

9. Quickly invert the MSA4 plate and drain the liquid onto an absorbent pad, then smack the plate down on a dry area of the pad, avoiding the liquid that was drained onto the pad.
10. Leave the plate uncovered and inverted on the tube rack for 1 h at room temperature to dry the pellet.

At this point the DNA can be resuspended as described below sealed and stored at $-25\text{ }^{\circ}\text{C}$ to $-15\text{ }^{\circ}\text{C}$ for no more than 24 h.

3.3.4 The Following Steps Describe the Resuspension of Precipitated DNA Samples

1. Before proceeding, preheat the Illumina Hybridization Oven to $48\text{ }^{\circ}\text{C}$ and the heat sealer for 20 min.
2. Thaw the RA1 reagent in a $20\text{--}25\text{ }^{\circ}\text{C}$ water bath. Gently mix to dissolve any crystals that might be present.
3. At the robot PC, select “MSA4|Resuspend MSA4”, and in the Basic Run Parameters change the value of number of MSA4 plates and number of DNA samples per plate to correspond with the number of samples being processed.
4. After placing the MSA4 plate on the robot bed, place a quarter reservoir in the reservoir frame, and add 4.5 mL of RA1 for 48 samples, or 9 mL for 96 samples.
5. Make sure the “Use Barcodes” checkbox is selected and then click “Run”.
6. Click “OK” in the message box, remove the MSA4 plate, and apply a foil seal by firmly holding the heat sealer block down for 3 s.
7. Immediately remove the plate from the heat sealer and forcefully roll the rubber plate sealer over the plate until 96 indentations become visible.
8. Place the sealed MSA4 plate in the Illumina Hybridization Oven to incubate for 1 h at $48\text{ }^{\circ}\text{C}$.
9. Vortex the plate at 1800 rpm for 1 min and pulse centrifuge at $280 \times g$.

At this point, the samples can be hybridized to the BeadChip, where it is safe to leave the RA1 at room temperature, or stored as the MSA4 plate at $-25\text{ }^{\circ}\text{C}$ to $-15\text{ }^{\circ}\text{C}$ for no more than 24 h. The plate can also be stored at $-80\text{ }^{\circ}\text{C}$ for longer than 24 h.

3.3.5 *The Following Steps Describe the Dispensation of the Fragmented, Resuspended DNA Samples onto BeadChips and Incubation of BeadChips in the Illumina Hybridization Oven for Sample Hybridization*

1. Preheat the heat block to 95 °C and the Illumina Hybridization Oven to 48 °C. Set the rocker speed to 5.
2. Make sure you have the correct Robot Tip Alignment Guide for the Infinium assay you are running. The barcode says Guide-B for the 12 × 1 HD BeadChip or Guide-E for the 8 × 1 HD Beadchip. Also, wash and dry the entire one-piece Robot Tip Alignment Guide.
3. Denature the samples in the MSA4 plate on the heat block at 95 °C for 20 min.
4. Remove the BeadChips from 2 to 8 °C storage, but leaving them in the original packaging until ready to use. During the 20-min incubation, prepare the Hyb Chambers.
5. Place the BeadChip Hyb Chamber gaskets into the BeadChip Hyb Chambers.
6. Dispense 400 µL PB2 reagent into the humidifying buffer reservoirs in the Hyb Chambers, and then immediately place the lid on the Hyb Chamber to prevent evaporation. At this point, the lid does not need to be locked.
7. Leave the closed Hyb Chambers on the bench at room temperature until the BeadChips are loaded with DNA sample. Load BeadChips into the Hyb Chamber within 1 h (*see Note 11*).
8. After the 20-min incubation, remove the MSA4 plate from the heat block and let sit at room temperature for 30 min.
9. Remove all BeadChips from packaging (*see Note 12*).
10. Place BeadChips into the Robot BeadChip Alignment Fixtures with the barcode end aligned to the ridges on the fixture.
11. Select “MSA4 Tasks | Hyb-Multi BC2” at the robot PC.
12. In the Basic Run Parameters pane, change the value for number of MSA4 plates and number of DNA samples per plate to correspond with the number of samples being processed.
13. Place the Robot BeadChip Alignment Fixtures onto the robot bed according to the bed map.
14. Pulse-centrifuge the plate to 280 × *g* and place it onto the robot bed. Remove the foil seal, make sure that the “Use Barcodes” checkbox is checked, and click “Run”.
15. Place the Robot Tip Alignment Guide on top of the Robot BeadChip Alignment Fixture with the Guide-B barcode upside down and facing away. Push both the Robot Tip Alignment Guide and Robot BeadChip Alignment Fixture to the upper left corner in its section of the robot bed.
16. At the robot PC, click “OK” to confirm that the Robot Tip Alignment Guide has been placed on top of the Robot

BeadChip alignment fixture. The robot scans the barcode on the Robot Tip Alignment Guide to confirm that the correct tip guide is being used. The robot dispenses sample to the BeadChips.

17. Click “OK” in the message box.
18. Carefully remove the Robot BeadChip alignment fixtures from the robot bed and visually inspect all sections of the BeadChips. Make sure that DNA sample covers all the sections of each bead stripe. Record any sections that are not completely covered.
19. Make sure that the Illumina Hybridization Oven is set to 48 °C.
20. Carefully remove each BeadChip from the Robot BeadChip alignment fixtures when the robot finishes (*see Note 13*).
21. Carefully place each BeadChip in a Hyb Chamber insert, orienting the barcode end so that it matches the barcode symbol on the insert.
22. Load the Hyb Chamber inserts containing loaded BeadChips inside the Illumina Hyb Chamber. Position the barcode over the ridges indicated on the Hyb Chamber.
23. In the Illumina LIMS left pane click “Infinium HD Methylation | Infinium Prepare Hyb Chamber”.
24. Scan the barcodes of the PB2 tubes and scan the BeadChip barcodes. Click “Verify” and “Save”.
25. Position the lid onto the Hyb Chamber by applying the back-side of the lid first, and then slowly bringing down the front end to avoid dislodging the Hyb Chamber inserts.
26. Close the clamps on both sides of the Hyb Chamber so that the lid is secure and even on the base and there are no gaps (*see Note 14*).
27. Place the Hyb Chamber in the 48 °C Illumina Hybridization Oven with the clamps on the left and right sides of the oven and the Illumina logo facing front.
28. Incubate at 48 °C for at least 16 h, but no more than 24 h.

*3.3.6 Proceed to Wash
BeadChips (Post-Amp)
After the Overnight
Incubation*

1. Add 330 mL 100% EtOH to the XC4 bottle, for a final volume of 350 mL. Each XC4 bottle has enough solution to process up to 24 BeadChips.
2. Shake the XC4 bottle vigorously to ensure complete resuspension. After it is resuspended, use XC4 at room temperature. The solution can be stored at 2 °C–8 °C for 2 weeks.
3. For optimal performance, wash and dry the Robot Tip Alignment Guides after every run.

4. Soak the tip guide inserts for 5 min in a 1% aqueous Alconox solution (1 part Alconox to 99 parts water) using a 400 mL Pyrex beaker (*see Note 15*).
5. After the soak in the 1% Alconox solution, thoroughly rinse the tip guides with DiH₂O at least three times to remove any residual detergent.
6. Dry the Robot Tip Alignment Guide using a Kimwipe or lint-free paper towels. Use a laboratory air gun to dry. Be sure to inspect the tip guide channels, including the top and bottom. Tip guides must be dry and free of any residual contaminants before next use.

*3.3.7 Prepare
the BeadChips
for the Staining Process*

1. Remove each Hyb Chamber from the Illumina Hybridization Oven and let cool for 30 min before opening.
2. Fill two wash dishes with PB1 (200 mL per dish) and fill the Multi-Sample BeadChip Alignment Fixture with 150 mL PB1.
3. Attach the wire handle to the rack and submerge the wash rack in the wash dish containing 200 mL PB1.
4. Remove the Hyb Chamber inserts from the Hyb Chambers, the BeadChip from the Hyb Chamber insert and finally, remove the cover seal from each BeadChip (*see Note 16*).
5. Using powder-free gloved hands, hold the BeadChip securely by the edges in one hand. Avoid contact with the sample inlets. Make sure that the barcode is facing up and closest to you, and that the top side of the BeadChip is angled slightly away from you. Remove the entire seal in a single, continuous motion. Start with a corner on the barcode end and pull with a continuous upward motion away from you and toward the opposite corner on the topside of the BeadChip. Do not touch the exposed arrays.
6. Immediately and carefully slide each BeadChip into the wash rack, making sure that the BeadChip is submerged in the PB1.
7. Repeat until all BeadChips (a maximum of eight) are transferred to the submerged wash rack.
8. After all BeadChips are in the wash rack, move the wash rack up and down for 1 min, breaking the surface of the PB1 with gentle, slow agitation.
9. Move the wash rack to the other wash dish containing clean PB1 and repeat **step 8**.
10. When you remove the BeadChips from the wash rack, inspect them for remaining residue.

11. For each additional set of eight BeadChips, assemble the flow-through chambers for the first eight BeadChips and repeat the wash steps (*see Note 17*).
12. If you plan to process more than four BeadChips, the 150 mL of PB1 used to fill the BeadChip alignment fixture can be reused for an additional set of four BeadChips (*see Note 18*).
13. For each BeadChip to be processed, place a black frame into the BeadChip alignment fixture prefilled with PB1.
14. Place each BeadChip to be processed into a black frame, aligning its barcode with the ridges stamped onto the alignment fixture (*see Note 19*).
15. Place a clear spacer onto the top of each BeadChip. Use the alignment fixture grooves to guide the spacers into proper position (*see Note 20*).
16. Place the alignment bar onto the alignment fixture. The groove in the alignment bar fits over the tab on the alignment fixture.
17. Place a clean glass back plate on top of the clear spacer covering each BeadChip. The plate reservoir is at the barcode end of the BeadChip, facing inward to create a reservoir against the BeadChip surface.
18. Attach the metal clamps to the flow-through chambers as follows: gently push the glass back plate against the alignment bar with one finger. Place the first metal clamp around the flow-through chamber so that the clamp is approximately 5 mm from the top edge. Place the second metal clamp around the flow-through chamber at the barcode end, approximately 5 mm from the reagent reservoir.
19. Using scissors, trim the ends of the clear plastic spacers from the flow-through chamber assembly. Slip scissors up over the barcode to trim the other end.
20. Immediately wash the Hyb Chamber reservoirs with DiH_2O and scrub them with a small cleaning brush, ensuring that no PB2 remains in the Hyb Chamber reservoir.

If you are using Illumina LIMS, in the Illumina LIMS left pane, click “Infinium HD Methylation|Wash BeadChip”. Scan the reagent barcodes and the BeadChip barcodes, click “Verify”, and then click “Save”. Illumina LIMS records the data and queues the BeadChips for the next step (*see Note 21*).

3.3.8 The Following Steps Describe the Washing of Unhybridized and Nonspecifically Hybridized DNA Sample from the BeadChips, Addition of Nucleotides to Extend the Primers to the DNA, Staining of the Primers, Disassembly of the Flow-through Chambers, and Coating of Beadchips for Protection

1. Thaw the RA1 reagent in a 20–25 °C water bath. Gently mix to dissolve any crystals that might be present.
2. Place all tubes (RA1, XC1, XC2, TEM, XC3, STM, ATM, PB1, XC4, Alconox Powder Detergent, 95% formamide/1 mM EDTA) in a rack in the order in which they will be used. If frozen, allow them to thaw to room temperature, and then gently invert the reagent tubes at least ten times to mix contents.
3. Ensure the water circulator is filled to the appropriate level. Turn on the water circulator and set it to 44 °C using the Circulator Manager in the automation control software.
4. Remove bubbles trapped in the Chamber Rack.
5. Test several locations on the Chamber Rack, using the Illumina Temperature Probe. All locations should be at 44 °C ± 0.5 °C. If the temperature on the probe is not within ±0.5 °C, contact Illumina Technical Support.

3.3.9 The Remaining Steps must Be Performed without Interruption

1. Slide the chamber rack into column 36 on the robot bed. Make sure that it is seated properly.
2. Select “XStain Tasks|XStain HD BeadChip” at the robot PC. In the Basic Run Parameters pane, enter the number of BeadChips.
3. If you plan on imaging the BeadChips immediately after the staining process, turn on the iScan or HiScan now to allow the lasers to stabilize.
4. Place a quarter reservoir in the reservoir frame, according to the robot bed, and add 15 mL of 95% formamide/1 mM EDTA to process 8 BeadChips, 17 mL to process 16 BeadChips, or 25 mL to process 24 BeadChips.
5. Place a half reservoir in the reservoir frame, according to the robot bed, and add 50 mL of CX3 to process 8 BeadChips, 100 mL to process 16 BeadChips, or 150 mL to process 24 BeadChips.
6. Place each reagent tube (XC1, XC2, TEM, STM, ATM) in the robot tube rack according to the bed map, and remove their caps.
7. When prompted, enter the stain temperature indicated on the STM tube. Do not load the BeadChips yet.
8. When the chamber rack reaches 44 °C, quickly place each flow-through chamber assembly into the first row of the chamber rack. Refer to the robot bed map for the correct layout.
9. Click “OK” at the robot PC. When the robot finishes, immediately remove the flow-through chambers from the chamber rack. Place horizontally on the lab bench at room temperature.

10. Pour 310 mL PB1 per 8 BeadChips into a wash dish.
11. Place the staining rack inside the wash dish.
12. For each BeadChip, use the dismantling tool to remove the two metal clamps from the flow-through chamber. Remove the glass back plate, the spacer, and then the BeadChip. Immediately place each BeadChip into the staining rack that is in the wash dish with the barcode facing away from you. Make sure all BeadChips are submerged.
13. Slowly move the staining rack up and down ten times, breaking the surface of the reagent.
14. Allow the BeadChips to soak for an additional 5 min.
15. Shake the XC4 bottle vigorously to ensure complete resuspension. If necessary, vortex until dissolved.
16. Pour 310 mL CX4 into a wash dish (*see Note 22*).
17. Move the BeadChip staining rack into the XC4 dish.
18. Slowly move the staining rack up and down ten times, breaking the surface of the reagent.
19. Allow the BeadChips to soak for an additional 5 min.
20. Lift the staining rack out of the solution and place it on a tube rack with the staining rack and BeadChips horizontal, barcodes facing up.
21. Remove the BeadChips from the staining rack with locking tweezers, working from top to bottom. Place each BeadChip on a tube rack to dry. Remove the rack handle if it facilitates removal of the BeadChips.
22. Dry the BeadChips in the vacuum desiccator for 50–55 min at 675 mm Hg (0.9 bar).
23. Make sure that the XC4 coating is dry before proceeding to the next step.
24. Clean the underside of each BeadChip with a ProStat EtOH wipe or Kimwipe soaked in EtOH (*see Note 23*).
25. Clean and store the glass back plates and Hyb Chamber components.
26. If you are using Illumina LIMS, in the Illumina LIMS left pane click “Infinium HD Methylation|Coat BC2”. Scan the reagent barcodes and BeadChip barcodes, and then click “Save”. Illumina LIMS records the data and queues the BeadChips for the next step.

At this point, you can either proceed to Image BeadChip (Post-Amp), or store the BeadChips in the Illumina BeadChip Slide Storage Box inside a vacuum desiccator at room temperature. Be sure to image the BeadChips within 72 h.

3.3.10 Image BeadChip (Post-Amp)

1. Follow the instructions in the iScan System User Guide or HiScan System User Guide to scan your BeadChips.
2. Use the Methylation NXT scan setting for your BeadChip.
3. Follow the instructions in the Decode File Client User Guide (11337856) available on the Illumina support website to download your DMAPs.
4. Use the GenomeStudio Methylation Module v1.8 Guide (11319130) to analyze your data. Use the Infinium > > Infinium HD setting to create a GenomeStudio project for your BeadChip. Analysis requires a sample sheet that describes the location of each sample.

3.4 Analysis of Illumina HumanMethylation450 BeadChip

3.4.1 Preprocessing

Raw IDAT files can be imported using the *minfi* package [18], followed by comprehensive evaluation of bisulfite conversion and subsequent array hybridization success rate. Filtering out problematic probes that failed to hybridize and are not represented by a minimum of three beads on the array is recommended before performing any downstream analysis. In practice, this step typically entails filtering out probes displaying a high detection p-value (e.g., >0.05). The number of nonspecific (cross-reactive) Infinium HumanMethylation450 probes ranges between 8.6 and 25% depending on the criteria used [19, 20]. This is particularly problematic, since methylation measurements from a nonspecific probe could represent signal from several genomic sites, alluding to false differential expression in further analyses. Therefore, it is of interest to remove probes that cross-hybridize to multiple genomic regions or are located on the sex chromosomes [21]. Finally, DNA methylation measurements can be confounded by the actual DNA sequence, particularly in independent case-control studies, and single nucleotide polymorphisms (SNPs) can be present within the remainder of the probe, although some studies suggest that DNA methylation measurements are not significantly affected by these [19].

3.4.2 Within-Array Normalization

The comprehensive nature of the 450k array is due to the use of two different types of chemical assays (~30% Infinium I and 70% Infinium II). However, the two probe types display a different dynamic range potentially leading to type II bias during analysis [22]. Complete Infinium HumanMethylation450 within-array normalization thus encompasses Infinium I/II-type bias correction, dye bias adjustment, and background correction. The two-assay design has led to the development of a plethora of within-array normalization algorithms. The R package *wateRmelon* [23] takes advantage of known methylation patterns that have been previously associated with genomic imprinting, X chromosome inactivation, and 65 SNPs present on the array to test methods of correction and normalization, while also providing access to 15 different

normalization methods. A noncomprehensive list of popular methods includes peak-based correction (PBC), Subset Quantile Normalization, Subset quantile for Within Array Normalization (SWAN) and Beta Mixture Quantile normalization (BMIQ) among others [22–26]. Color bias adjustment methods based on smooth quantile and shift-and-scaling normalization and Genome Studio are implemented in *lumi* and *methylumi* R packages [27]. Finally, some of the freely available background correction methods include simple background subtraction, mixture models that model signal intensity and background noise separately (ENmix), and convolution models implemented in *methylumi* (normal-exponential convolution using out-of-band probes) [28].

3.4.3 Between-Array Normalization and Batch Effects

Between-array normalization is essential to remove sources of non-biological variation related to external parameters, including reagent concentrations and temperature, unequal quantities of starting material, and detection efficiencies. Simple, yet effective options, derived from processing methods initially developed for gene expression arrays, are the different versions of the quantile normalization. Other between-array methods recently developed include shift and scaling normalization of the *lumi* package, local regression-based approaches [29], and unsupervised functional normalization [30]. Batch and side effects can generate artifacts on methylation measurements at the global level that could be partially alleviated due to between-array normalization. However, global between-array methods may be inefficient in settings where batch effects affect only a subset of probes. Popular methods for batch effect correction are the supervised methods *ComBat* [31], independent surrogate variable analysis (ISVA) [32], and surrogate variable analysis (SVA) [33] implemented in the *RnBeads* package.

3.4.4 Differential Methylation Analysis

After the preprocessing and normalization steps, the calculation of differentially methylated positions (DMPs) can be performed. In epigenome-wide association studies, one might want to consider correction for cell heterogeneity first [34]. The β -value is defined as the ratio of the methylated signal over the total signal used to express the degree of methylation obtained with Infinium array. The more heteroscedasticity resistant statistic, the M-value, is the log-transformed ratio of the methylated over the unmethylated signal. Statistical tests for simple, independent two-sample designs consisting of t-test or Mann-Whitney test perform best on M-values. In addition to single CpG analysis, one might also want to identify differentially methylated regions (DMRs). The intuition behind measurement of DMRs derives from the idea that probes lying within the promoter of the same gene, or in a window of a given size, should exhibit the same methylation patterns. Depending on the genomic region, the *minfi* package implements two

different functions for estimating DMRs [18]: *bumphunter* focuses on short range (1–2 kb) methylation changes (around gene promoters), while *cpGCollapse* covers long-range changes as represented by the 170,000 open sea probes on the 450k array. Another freely available DMR method includes *champ.lasso* function, implemented in Chip Analysis Methylation Pipeline (*ChAMP*), which uses a dynamic window based on genomic features to capture DMRs [35]. Other recently developed DMR calling methods include *DMRcate*, *FastDMA*, and *COHCAP* that identifies CpG islands showing consistent methylation pattern among CpG sites. The *COHCAP* package also allows for gene expression and DNA methylation data integration for identification of CpG islands that potentially regulate downstream gene expression [36].

4 Notes

1. We recommend that this solution be prepared in a plastic container, as the reagent may absorb to glass surfaces. The working solution should also be protected from light, as the Quant-iT PicoGreen reagent is photodegradable. For best results, prepare this solution within a few hours of using.
2. To ensure that sample readings remain in the detection range of the fluorometer, the instrument's gain should be set so that the sample containing the highest DNA concentration yields a fluorescence intensity near the instrument's maximum range.
3. A higher dilution of the experimental sample may diminish the interfering effect of certain contaminants, but extremely small sample volumes should be avoided because they are difficult to pipet accurately.
4. To minimize photobleaching effects, keep the time for fluorescence measurement constant for all samples.
5. Minimize the exposure to light of the CT Conversion Reagent, as it is light-sensitive. For best results, the CT Conversion Reagent should be used immediately following preparation. The prepared reagent can be stored overnight at room temperature, 1 week at 4 °C, or up to 1 month at –20 °C. Stored CT Conversion Reagent must be warmed to 37 °C and vortexed prior to use.
6. 1–4 µL of eluted DNA is recommended for each PCR. The elution volume can be more than 10 µL, but small elution volumes are preferred for more concentrated DNA.
7. If you do not already have a WG#-BCD plate, add the bisulfite-converted DNA to either a Midi plate: 20 µL to each WG#-BCD plate well or to a TCY plate: 10 µL to each WG#-BCD plate well. Apply a barcode label to the new WG#-BCD plate.

8. When you remove a cap mat, set it aside, upside down, in a safe location for use later in the protocol. When you place the cap mat back on the plate, be sure to match it to its original plate and in the correct orientation.
9. If you are using Illumina LIMS, you cannot change the number of DNA samples on this screen. However, it will process the correct number of samples.
10. Set the centrifuge to 4 °C in preparation for the next centrifuge step.
11. You can also prepare the Hyb Chambers later, during the 30-min cool down.
12. When handling the BeadChip, avoid touching the beadstripe area and sample inlets.
13. For optimal performance, take care to keep the Hyb Chamber inserts containing BeadChips steady and level when lifting or moving. Avoid shaking and always keep parallel to the lab bench. Do not hold by the sides near the sample inlets.
14. Keep the Hyb Chamber steady and level when moving it or transferring it to the Illumina Hybridization Oven.
15. Do not use bleach or ethanol to clean the tip guide inserts.
16. Remove the cover seal over an absorbent cloth or paper towels, preferably in the hood to make sure that no solution splatters on you.
17. Confirm that you are using the correct Infinium standard glass back plates and spacers before assembling the flow-through chambers. Refer to Fig. 2 for the correct flow-through chamber components.
18. Use 150 mL of PBI for every additional set of eight BeadChips.
19. Inspect the surface of each BeadChip for residue left by the seal. Use a pipette tip to remove any residue under buffer and be careful not to scratch the bead area.
20. Be sure to use the clear plastic spacers, not the white ones.
21. Place all assembled flow-through chambers on the lab bench in a horizontal position while you perform the preparation steps for the XStain BeadChip. Do not place the flow-through chambers in the chamber rack until the preparation is complete.
22. Do not let the XC4 sit for longer than 10 min.
23. Do not touch the stripes with the wipe or allow EtOH to drip onto the stripes.

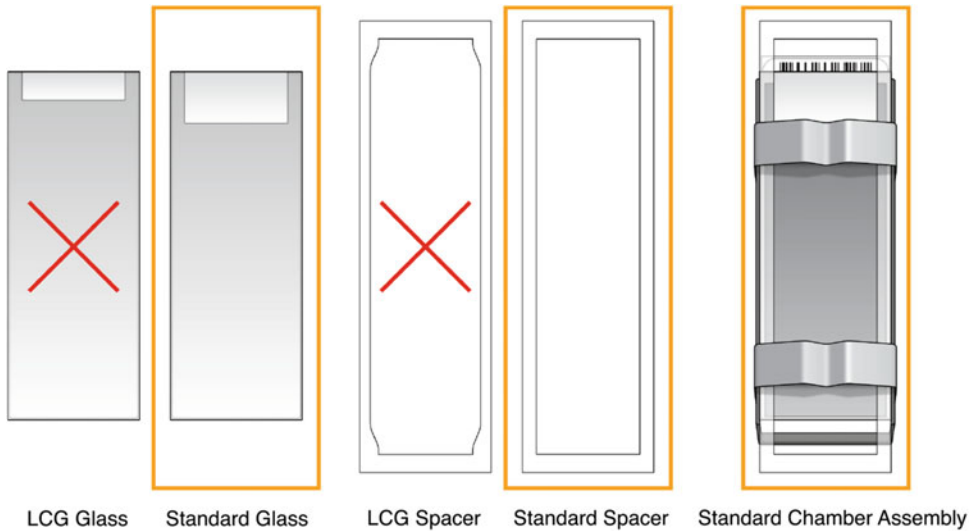


Fig. 2 Assembly of flow-through chambers. This figure was adapted from the Illumina Infinium HumanMethylation450 BeadChip Kit protocol

Acknowledgments

We thank Diego Portillo Santos for editing and generating the figures.

References

- Bird A, Taggart M, Frommer M, Miller OJ, Macleod D (1985) A fraction of the mouse genome that is derived from islands of non-methylated, CpG-rich DNA. *Cell* 40(1):91–99
- Holliday R, Pugh JE (1975) DNA modification mechanisms and gene activity during development. *Science* 187(4173):226–232
- Bird A (2007) Perceptions of epigenetics. *Nature* 447(7143):396–398. <https://doi.org/10.1038/nature05913>
- Dayeh T, Volkov P, Salo S, Hall E, Nilsson E, Olsson AH et al (2014) Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. *PLoS Genet* 10(3):e1004160. <https://doi.org/10.1371/journal.pgen.1004160>. PubMed PMID: 24603685; PubMed Central PMCID: PMC3945174.
- Kulis M, Esteller M (2010) DNA methylation and cancer. *Adv Genet* 70:27–56. <https://doi.org/10.1016/B978-0-12-380866-0.60002-2>
- Li E (2002) Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet* 3(9):662–673. <https://doi.org/10.1038/nrg887>
- Lu H, Liu X, Deng Y, Qing H (2013) DNA methylation, a hand behind neurodegenerative diseases. *Front Aging Neurosci* 5:85. <https://doi.org/10.3389/fnagi.2013.00085>. PubMed PMID: 24367332; PubMed Central PMCID: PMC3851782.
- Zhong J, Agha G, Baccarelli AA (2016) The role of DNA methylation in cardiovascular risk and disease: methodological aspects, study design, and data analysis for epidemiological studies. *Circ Res* 118(1):119–131. <https://doi.org/10.1161/CIRCRESAHA.115.305206>. PubMed PMID: 26837743; PubMed Central PMCID: PMC4743554
- Goll MG, Bestor TH (2005) Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* 74:481–514. <https://doi.org/10.1146/annurev.biochem.74.010904.153721>
- Feng S, Jacobsen SE, Reik W (2010) Epigenetic reprogramming in plant and animal development. *Science* 330(6004):622–627.

- <https://doi.org/10.1126/science.1190614>. PubMed PMID: 21030646; PubMed Central PMCID: PMCPMC2989926
11. Deaton AM, Bird A (2011) CpG islands and the regulation of transcription. *Genes Dev* 25 (10):1010–1022. <https://doi.org/10.1101/gad.2037511>. Epub 2011/05/18. doi. PubMed PMID: 21576262; PubMed Central PMCID: PMC3093116.
 12. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL et al (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* 28 (10):1097–1105. <https://doi.org/10.1038/nbt.1682>. PubMed PMID: 20852635; PubMed Central PMCID: PMCPMC2955169
 13. Kurdyukov S, Bullock M (2016) DNA methylation analysis: choosing the right method. *Biology (Basel)* 5(1). <https://doi.org/10.3390/biology5010003>. PubMed PMID: 26751487; PubMed Central PMCID: PMCPMC4810160
 14. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW et al (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* 89 (5):1827–1831. PubMed PMID: 1542678; PubMed Central PMCID: PMCPMC48546
 15. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL et al (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* 37(8):853–862. <https://doi.org/10.1038/ng1598>
 16. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD et al (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452(7184):215–219. <https://doi.org/10.1038/nature06745>. PubMed PMID: 18278030; PubMed Central PMCID: PMCPMC2377394
 17. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM et al (2011) High density DNA methylation array with single CpG site resolution. *Genomics* 98(4):288–295. <https://doi.org/10.1016/j.ygeno.2011.07.007>
 18. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD et al (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30(10):1363–1369. <https://doi.org/10.1093/bioinformatics/btu049>. PubMed PMID: 24478339; PubMed Central PMCID: PMCPMC4016708
 19. Price ME, Cotton AM, Lam LL, Farre P, Emberly E, Brown CJ et al (2013) Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* 6(1):4. <https://doi.org/10.1186/1756-8935-6-4>. PubMed PMID: 23452981; PubMed Central PMCID: PMCPMC3740789
 20. Zhang X, Mu W, Zhang W (2012) On the analysis of the illumina 450k array data: probes ambiguously mapped to the human genome. *Front Genet* 3:73. <https://doi.org/10.3389/fgene.2012.00073>. PubMed PMID: 22586432; PubMed Central PMCID: PMCPMC3343275
 21. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW et al (2013) Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 8(2):203–209. <https://doi.org/10.4161/epi.23470>. PubMed PMID: 23314698; PubMed Central PMCID: PMCPMC3592906
 22. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D et al (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 29(2):189–196. <https://doi.org/10.1093/bioinformatics/bts680>. PubMed PMID: 23175756; PubMed Central PMCID: PMCPMC3546795
 23. Pidsley R, CC YW, Volta M, Lunnon K, Mill J, Schalkwyk LC (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* 14:293. <https://doi.org/10.1186/1471-2164-14-293>. PubMed PMID: 23631413; PubMed Central PMCID: PMCPMC3769145
 24. Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, Fuks F (2014) A comprehensive overview of Infinium HumanMethylation450 data processing. *Brief Bioinform* 15(6):929–941. <https://doi.org/10.1093/bib/bbt054>. PubMed PMID: 23990268; PubMed Central PMCID: PMCPMC4239800
 25. Maksimovic J, Gordon L, Oshlack A (2012) SWAN: subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol* 13(6):R44. <https://doi.org/10.1186/gb-2012-13-6-r44>. PubMed PMID: 22703947; PubMed Central PMCID: PMCPMC3446316

26. Touleimat N, Tost J (2012) Complete pipeline for Infinium((R)) human methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* 4(3):325–341. <https://doi.org/10.2217/epi.12.21>
27. Du P, Kibbe WA, Lin SM (2008) lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 24(13):1547–1548. <https://doi.org/10.1093/bioinformatics/btn224>
28. Xu Z, Niu L, Li L, Taylor JA (2016) ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Res* 44(3):e20. <https://doi.org/10.1093/nar/gkv907>. PubMed PMID: 26384415; PubMed Central PMCID: PMC4756845
29. Heiss JA, Brenner H (2015) Between-array normalization for 450K data. *Front Genet* 6:92. <https://doi.org/10.3389/fgene.2015.00092>. PubMed PMID: 25806048; PubMed Central PMCID: PMC4354407
30. Fortin JP, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ et al (2014) Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol* 15(11):503. <https://doi.org/10.1186/s13059-014-0503-2>. PubMed PMID: 25599564; PubMed Central PMCID: PMC4283580
31. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8(1):118–127. <https://doi.org/10.1093/biostatistics/kxj037>. PubMed PMID: 16632515
32. Teschendorff AE, Zhuang J, Widschwendter M (2011) Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* 27(11):1496–1505. <https://doi.org/10.1093/bioinformatics/btr171>
33. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 3(9):e161–e135. <https://doi.org/10.1371/journal.pgen.0030161>. Epub 2007/10/03. PubMed PMID: 17907809; PubMed Central PMCID: PMC1994707
34. Jaffe AE, Irizarry RA (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol* 15(2):R31. <https://doi.org/10.1186/gb-2014-15-2-r31>. PubMed PMID: 24495553; PubMed Central PMCID: PMC4053810
35. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK et al (2014) ChAMP: 450k Chip analysis methylation pipeline. *Bioinformatics* 30(3):428–430. <https://doi.org/10.1093/bioinformatics/btt684>. PubMed PMID: 24336642; PubMed Central PMCID: PMC43904520
36. Warden CD, Lee H, Tompkins JD, Li X, Wang C, Riggs AD et al (2013) COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. *Nucleic Acids Res* 41(11):e117. <https://doi.org/10.1093/nar/gkt242>. PubMed PMID: 23598999; PubMed Central PMCID: PMC43675470

Part III

Functional Characterization of Susceptibility Alleles and Loci

Chapter 14

miRNA Quantification Method Using Quantitative Polymerase Chain Reaction in Conjunction with C_q Method

Fatjon Leti and Johanna K. DiStefano

Abstract

MicroRNAs are small noncoding RNAs that function to regulate gene expression. In general, miRNAs are posttranscriptional regulators that imperfectly bind to the 3' untranslated region (3'UTR) of target mRNAs bearing complementary sequences, and target more than half of all protein-coding genes in the human genome. The dysregulation of miRNA expression and activity has been linked with numerous diseases, including cancer, cardiovascular diseases, neurodegenerative disorders, and diabetes. To better understand the relationship between miRNAs and human disease, a variety of techniques have been used to measure and validate miRNA expression in many cells, tissues, body fluids, and organs. For many years, quantitative polymerase chain reaction (qPCR) has been the gold standard for measuring relative gene expression, and is now also widely used to assess miRNA abundance. In this chapter, we describe a quick protocol for miRNA extraction, reverse transcription, qPCR, and data analysis.

Key words miRNA, Reverse transcription, RT-PCR, qRT-PCR, qPCR, C_q method, Delta-delta C_q

1 Introduction

MicroRNAs (miRNAs) are single-stranded RNAs (approximately 20–24 nucleotides in length), endogenously expressed in most eukaryotic cells. Dysregulation of miRNAs has been associated with many types of cancers, viral infections, cardiovascular diseases, neurodegenerative disorders, and diabetes [1]. In animals, miRNAs function to suppress translation or initiate degradation of target mRNAs by binding to the complementary sequence in the 3' untranslated region [2]. Imperfect binding between a given miRNA and complementary sequences allows a single miRNA to target a multitude or different genes. Computational analyses reveal that 60% of mRNAs are potential miRNA targets [3]. There are many free publically available databases that predict potential mRNA targets for a miRNA of interest [4–7]. Although such platforms are very helpful, they often yield many false positives; thus, experimental validation and quantification of a particular miRNA is

necessary to confirm predicted miRNA-mRNA interactions. Typical validation and quantification strategies involve luciferase reporter gene assay, fluorescence-based quantitative real-time PCR (qPCR), and less commonly, northern blotting. In this chapter, we focus on the qPCR assay, which is recognized as the gold standard for expression analysis because it provides the widest dynamic range, the lowest quantification limits, and the least biased results compared to next generation sequencing and microarrays [8–10].

qPCR is very similar to traditional PCR, the only difference being that the product is measured after each round in real time, rather than at the endpoint of amplification. This allows quantification of transcript abundance between controlled and experimental conditions. Analysis of qPCR results can use absolute or relative values. In absolute quantitation, the unknown quantity is compared to a standard curve comprised of known values, while the relative quantification compares miRNA levels between two different conditions. Here, we present the $\Delta\Delta C_q$ method [11], which requires a reference control to normalize the miRNA target, as an approach to measure relative quantification. Selecting a constant reference control is a very important consideration in the experimental design as it is an indicator of RNA extraction batch effect, as well as reverse transcription [11].

In this chapter, we describe a protocol for miRNA extraction, cDNA template synthesis via reverse transcription, and qPCR amplification, in conjunction with a light cycler (Fig. 1). We also show a sample calculation using a miRNA (miR-182) that we found to be differentially expressed between normal and fibrotic liver samples in individuals with nonalcoholic fatty liver disease [12].

2 Materials

2.1 miRNA Extraction

1. Cell culture (*see Note 1*).
2. miRvana miRNA Isolation Kit (Thermo Fisher Scientific; Waltham, MA).
3. Acid-phenol–chloroform.
4. RNase-Free 1.5 polypropylene microfuge tubes.
5. Molecular grade 100% ethanol.
6. $1\times$ PBS.
7. Cell scrapers.
8. NanoDrop™ Spectrophotometer (Thermo Fisher Scientific).
9. RNaseZap.

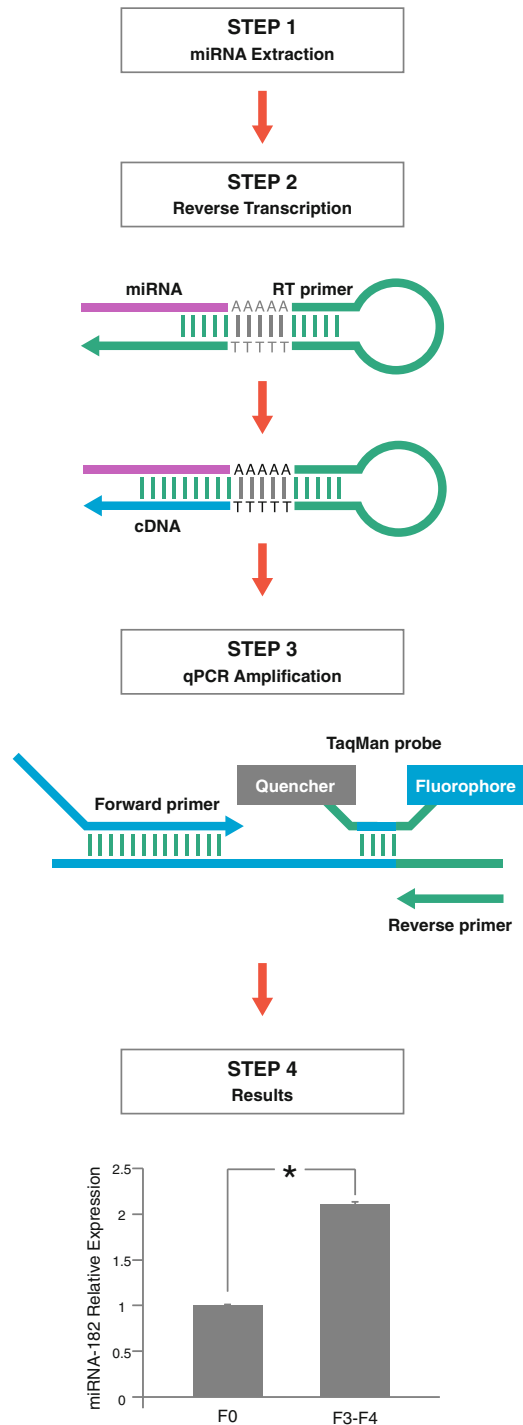


Fig. 1 Schematic representation of workflow to measure relative miRNA expression using the C_q method

2.2 Reverse Transcription

1. TaqMan[®] MicroRNA Reverse Transcription Kit (Thermo Fisher Scientific).
2. RNase- and DNase-free 0.2 mL PCR tubes or plates.

2.3 Quantitative PCR

1. TaqMan[®] MicroRNA assays (Thermo Fisher Scientific) (*see Note 2*).
2. 2× TaqMan[®] Universal PCR Master Mix, no AmpErase[®] UNG (Thermo Fisher Scientific) (*see Note 3*).

2.4 Data Analysis

1. Microsoft Excel or comparable program with basic statistical functions.

3 Methods
3.1 miRNA Extraction

1. Wipe work surface and pipettes using RNaseZap.
2. Add 21 mL of 100% ethanol to Wash Solution 1, provided with the miRvana miRNA Isolation kit. Mix well.
3. Add 40 mL of 100% ethanol to Wash Solution 2/3 and mix well.
4. For adherent cells aspirate and discard culture medium.
5. Wash cells with PBS and place plate on ice.
6. Aspirate PBS from the cells.
7. Add RNAlater if cells will be stored prior to RNA extraction; otherwise, proceed to **step 8**.
8. Lyse cells directly on the plate culture by adding 300–600 μ L of Lysis/Binding Solution (*see Note 2*).
9. Collect the cells with a cell scraper and transfer the lysate to a fresh tube.
10. Vortex vigorously until a homogeneous lysate is obtained.
11. Add 1/10 volume of miRNA Homogenate Additive to the cell lysate and vortex well (*see Note 3*).
12. Incubate the tube on ice for 10 min.
13. Add acid-phenol–chloroform mixture in a volume equivalent to that of the lysate in **step 8** (*see Note 4*).
14. Mix well by vortexing the tube for 30–60 s.
15. Centrifuge the mixture for 5 min at 10,000 $\times g$, to separate the aqueous and organic phases. If the interphase is not compact, repeat **step 15** (*see Note 5*).
16. Transfer the upper aqueous phase to a fresh tube (do not disturb the lower phase), and measure the volume recovered.

17. Add 1/3 volume of 100% ethanol to the aqueous phase retrieved from **step 16**. Mix samples by vortexing or inverting the tube (*see Note 6*).
18. Place one filter cartridge from the kit into a collection tube for each sample.
19. Pipette up to 700 μL of the lysate/ethanol mixture.
20. Centrifuge for ~ 15 s at $10,000 \times g$, and collect the filtrate.
21. If the volume of lysate/ethanol mixture is greater than 700 μL , repeat **steps 19** and **20**, and combine filtrates of each sample. The filter cartridge contains an RNA fraction that is depleted of small RNAs. *See Note 7* for steps on how to recover this fraction.
22. Add 2/3 volume of room temperature 100% ethanol to the filtrate from **step 21**, and mix well (*see Note 8*).
23. Transfer 700 μL of filtrate/ethanol mix to a new filter cartridge.
24. Centrifuge for ~ 15 s at $10,000 \times g$, and discard the eluate.
25. Repeat until all of the filtrate/ethanol mixture is run through the filter.
26. Add 700 μL of miRNA Wash Solution 1 to the filter cartridge and centrifuge at $10,000 \times g$ for 10 s. Discard eluate.
27. Apply 500 μL Wash Solution 2/3 and centrifuge at $10,000 \times g$ for 10 s. Discard eluate.
28. Repeat **step 27**.
29. Place filter cartridge in a new collection tube, and spin for 1 min at $10,000 \times g$ to remove residual fluid.
30. Transfer filter cartridge to a new collection tube, and add 100 μL of preheated 95°C nuclease-free water to the center of the filter.
31. Spin for 30 s at $\sim 16,000 \times g$ to recover RNA.
32. Proceed with the RNA quantitation and quality assessment using the NanoDrop Spectrophotometer (*see Note 9*).
33. Freeze samples at -80°C or continue with reverse transcription.

3.2 Reverse Transcription

1. Thaw the components of the TaqMan[®] MicroRNA Reverse Transcription Kit on ice.
2. Dilute primers to $5\times$ using $0.1\times$ TE buffer. Keep tubes on ice.
3. Prepare the Master Mix in a 0.2 mL PCR tube or 96-well plate following the volumes listed in Table 1. Prepare 15% excess volume to account for volume loss during pipetting.

Table 1
Reaction volumes for reverse transcription

Component	Vol/ 15 μ L	Vol/15 μ L + 15%
100 mM dNTPs (with dTTP)	0.15	0.17
MultiScribe reverse transcriptase, 50 U/ μ L	1.00	1.15
10 \times Reverse transcription buffer	1.50	1.73
5 \times RT primer	3.00	3.45
RNA sample (1–10 ng)	5.00	5.75
RNase inhibitor, 20 U/ μ L	0.19	0.22
Nuclease-free water	4.16	4.78
Total volume	15.00	17.25

Table 2
Thermocycler parameters for RT

Step	Time (min)	Temperature ($^{\circ}$ C)
Hold	30	16
Hold	30	42
Hold	5	85
Hold	∞	4

4. Mix gently by pipetting up and down a few times, and then centrifuge to bring the solution to the bottom of the plate. Keep plate on ice.
5. To the 7 μ L of Master Mix, add 5 μ L of RNA (1–10 ng per reaction) extracted from cells, and 3 μ L of primer from each assay.
6. Seal the plate and mix by inverting. Centrifuge to bring the solution to the bottom of the plate. Keep plate on ice until ready to load the thermocycler.
7. Use the parameters listed in Table 2 to program the thermocycler.
8. Load the plate on to the thermocycler, and start the RT program.
9. Once the run is completed, continue with the qPCR amplification or store the RT reactions at -15 to -25 $^{\circ}$ C.

3.3 qPCR Amplification

1. On ice, thaw TaqMan Assay (20×) and cDNA samples prepared above.
2. Resuspend by vortexing, followed by a brief centrifugation.
3. Gently swirl TaqMan 2× Universal PCR Master Mix, no AmpErase UNG.
4. Calculate the number of reactions, including the miRNA(s) of interest, endogenous control assay(s), and a no-template control (NTC) for each assay on the plate.
5. In a 1.5 mL sterile microcentrifuge tube, combine the reagents listed in Table 3. Prepare 15% excess volume (*see Note 10*).
6. Mix by inverting the tube several times, and centrifuge briefly.
7. Pipette 20 μL into each of the three wells on a 96- or 384-well plate.
8. Seal plate with an optically clear cover, and centrifuge briefly.
9. Load plate on to real-time instrument (*see Note 11*).
10. Start run using the parameters listed in Table 4.

3.4 Data Analysis

1. The *QuantStudio 7 Flex* Real-Time PCR System automatically determines and calculates baseline and threshold levels (*see Note 11*).
2. Import C_q values into Microsoft Excel.

Table 3
Volumes per 15 μL single reaction used for qPCR amplification

Component	Vol/15 μL (μL)	Vol/15 μL + 15% (μL)
TaqMan small RNA assay (20×)	1.00	1.15
Product from RT reaction*	1.33	1.53
TaqMan 2× universal PCR master mix, no AmpErase UNG	10.00	11.50
Nuclease-free water	7.67	8.82
Total volume	20.00	23.00

Table 4
Sample calculations for relative quantification

Phenotype	miRNA	C _{q1}	C _{q2}	C _{q3}	Average C _q	ΔC _q	ΔΔC _q	2 ^{-ΔΔC_q}
Control	miR-182	27.3	27.2	27.1	27.2 ± 0.1	6.8	0	1 ± 0.01
Control	U6	20.3	20.3	20.5	20.4 ± 0.1			
Case	miR-182	26.1	25.9	25.9	26.0 ± 0.1	5.7	-1.1	2.1 ± 0.02
Case	U6	20.2	20.1	20.4	20.2 ± 0.2			

3. If the difference among triplicates for each sample is greater than $0.5 C_q$, discard these values. Differences should be $<0.5 C_q$.
4. Average triplicates of each conditions, and calculate $\Delta C_q = [C_q \text{ miRNA of interest} - C_q \text{ reference miRNA}]$
5. Calculate $\Delta\Delta C_q$ where $\Delta\Delta C_q = [C_q \text{ miRNA sample A} - C_q \text{ miRNA sample B}]$.
6. Calculate relative fold-change using $2^{-\Delta\Delta C_q}$.
7. Refer to Table 4 for sample calculation.

4 Notes

1. This protocol applies only to adherent cells. Refer to the manufacturer's protocol for extraction of small RNAs from cells grown in suspension, fresh and frozen tissue samples, and yeast or bacterial cultures.
2. 300 μL of the Lysis/Binding Solution is recommended for small cell numbers (up to hundreds), while 600 μL is recommended for larger numbers of cells.
3. If the lysate volume is 300 μL , add 30 μL of miRNA Homogenate Additive.
4. If the lysate volume is 300 μL , add 300 μL of acid-phenol-chloroform mixture.
5. If the interphase is not compact after the second spin, adjust the speed to $16,000 \times g$ for 10 min.
6. If the aqueous phase was 300 μL , add 100 μL 100% ethanol.
7. See steps 26–31 in Subheading 3.1. RNA Extraction.
8. If the filtrate volume recovered is 400 μL , add 266 μL of 100% ethanol.
9. The simplest method for quantitating RNA to measure absorbance using the NanoDrop Spectrophotometer, for example. Good quality RNA has a A_{260}/A_{280} ratio of 1.8–2.1. The Quant-iT RiboGreen RNA Reagent and Kit is another available assay widely used to quantitate RNA.
10. The recommended reaction volume by the manufacturer is 20 μL . We have successfully quantified miRNAs by adjusting the volume reaction to 10 μL .
11. Please refer to your qPCR system for instructions on how to configure the experiment/plate document and run the PCR-plates.

Acknowledgments

We thank Diego Portillo Santos for generating the figure in this chapter.

References

1. Li Y, Kowdley KV (2012) MicroRNAs in common human diseases. *Genomics Proteomics Bioinformatics* 10:246–253
2. Du T, Zamore PD (2007) Beginning to understand microRNA function. *Cell Res* 17:661–663
3. Friedman RC, Farh KK, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 19:92–105
4. Dweep H, Sticht C, Pandey P, Gretz N (2011) miRWalk--database: prediction of possible miRNA binding sites by “walking” the genes of three genomes. *J Biomed Inform* 44:839–847
5. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS (2003) MicroRNA targets in drosophila. *Genome Biol* 5:R1
6. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 34: D140–D144
7. Hsu HH, Hoffmann S, Endlich N, Velic A, Schwab A, Weide T, Schlatter E, Pavenstadt H (2008) Mechanisms of angiotensin II signaling on cytoskeleton of podocytes. *J Mol Med* 86:1379–1394
8. Li P, Piao Y, Shon HS, Ryu KH (2015) Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics* 16:347
9. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2(-\Delta\Delta C(T))$ method. *Methods* 25:402–408
10. VanGuilder HD, Vrana KE, Freeman WM (2008) Twenty-five years of quantitative PCR for gene expression analysis. *BioTechniques* 44:619–626
11. Huggett J, Dheda K, Bustin S, Zumla A (2005) Real-time RT-PCR normalisation; strategies and considerations. *Genes Immun* 6:279–284
12. Leti F, Malenica I, Doshi M, Courtright A, Van Keuren-Jensen K, Legendre C, Still CD, Gerhard GS, DiStefano JK (2015) High-throughput sequencing reveals altered expression of hepatic microRNAs in nonalcoholic fatty liver disease-related fibrosis. *Transl Res* 166:304–314

Chapter 15

Primary Airway Epithelial Cell Gene Editing Using CRISPR-Cas9

Jamie L. Everman, Cydney Rios, and Max A. Seibold

Abstract

The adaptation of the clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR associated endonuclease 9 (CRISPR-Cas9) machinery from prokaryotic organisms has resulted in a gene editing system that is highly versatile, easily constructed, and can be leveraged to generate human cells knocked out (KO) for a specific gene. While standard transfection techniques can be used for the introduction of CRISPR-Cas9 expression cassettes to many cell types, delivery by this method is not efficient in many primary cell types, including primary human airway epithelial cells (AECs). More efficient delivery in AECs can be achieved through lentiviral-mediated transduction, allowing the CRISPR-Cas9 system to be integrated into the genome of the cell, resulting in stable expression of the nuclease machinery and increasing editing rates. In parallel, advancements have been made in the culture, expansion, selection, and differentiation of AECs, which allow the robust generation of a bulk edited AEC population from transduced cells. Applying these methods, we detail here our latest protocol to generate mucociliary epithelial cultures knocked out for a specific gene from donor-isolated primary human basal airway epithelial cells. This protocol includes methods to: (1) design and generate lentivirus which targets a specific gene for KO with CRISPR-Cas9 machinery, (2) efficiently transduce AECs, (3) culture and select for a bulk edited AEC population, (4) molecularly screen AECs for Cas9 cutting and specific sequence edits, and (5) further expand and differentiate edited cells to a mucociliary airway epithelial culture. The AEC knockouts generated using this protocol provide an excellent primary cell model system with which to characterize the function of genes involved in airway dysfunction and disease.

Key words CRISPR, Gene editing, Lentivirus, Airway epithelial cells, Primary cells, Gene knockout

1 Introduction

Research into the etiology of airway and other complex lung diseases has been greatly aided by agnostic, genome-wide genetic and genomic studies. These studies have generated lists of candidate genes likely involved in disease pathogenesis. However, many of the identified genes are novel with no known function or apparent mechanism for how their perturbation could contribute to development or exacerbation of human disease. Gene loss-of-function studies provide a powerful, causal experimental design that can be

employed to determine how a biological system responds to absence of a particular gene. Gene knockout studies of model organisms, such as mice, have been used for over 25 years to understand gene function at a whole organism level. Although these model organism studies are informative, the many differences between mice and humans in anatomy, physiology, and genetics do not obviate the need to examine the normal function and dysfunction of these genes in primary human cell types.

Gene knockouts in human cells have been difficult to generate due to poor rates of genome modification or editing using homologous recombination techniques employed in other organisms. Instead gene loss-of-function studies in human cells have focused on degradation of gene mRNA products through RNA interference (RNAi) [1, 2]. Such techniques utilize short interfering RNAs (siRNA) to bind to complementary mRNA transcripts, which are then recognized and cleaved by the Dicer endonuclease and degraded by the RNA-Induced Silencing Complex (RISC) [3, 4]. These experiments result in knockdown of protein expression for the targeted mRNA transcript, while maintaining the genetic code from which the mRNA was transcribed. The pitfalls of RNAi include the transient and incomplete nature of the RNA interference, and the extensive optimization necessary to determine both the initiation and duration of the knockdown at both the mRNA and protein levels [5].

Molecular gene editing techniques have since been developed that greatly enhance the efficiency of DNA level gene knockout generation in human cells, without the need for homologous recombination. Improved gene editing was first accomplished through advances in the design and construction of DNA site-specific binding proteins. Specifically, modifications of naturally occurring zinc finger (ZF) and transcription activator-like effector (TALE) protein domains were arrayed to construct recombinant proteins capable of recognizing and binding to specific DNA sequences [6, 7]. These DNA target site-specific binding proteins were fused to nuclease domains, allowing the generation of a site-specific double-stranded break (DSB) in the targeted DNA sequence. Specifically, the fusion of ZF and TALE proteins to nucleases (N) forms the basis for various ZFN and TALEN gene editing systems, respectively [8, 9]. Repair of these DSBs can occur by one of two different cellular mechanisms and allows for editing of the gene sequence. The homology-directed repair mechanism can occur if a DNA fragment homologous to the cut site is provided to the cells. Alternatively, if no homologous template is present, the cell can repair the break by ligating the two DNA ends together using nonhomologous end joining (NHEJ). This process is error-prone and results in the generation of random insertions and deletions (indels) at the break site with a high frequency. When the programmed nuclease is targeted to a gene coding exon the

resultant indels are likely to produce a codon frameshift and permanent loss of functional protein expression.

Although both ZFN and TALEN gene editing systems have been and continue to be successfully applied to a variety of cell types, they do have a significant limitation. Namely, since the DNA site specificity lies in the protein sequence, they require the difficult generation (in both a design and molecular construction sense) of a different recombinant protein for each targeted DNA site. This limitation was resolved by the development of the Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR-associated protein 9 (CRISPR-Cas9) gene editing system, which is guided to a DNA target sequence by a small complementary RNA sequence rather than peptide sequences. The CRISPR-Cas9 system is derived from components of a prokaryotic type II adaptive immune response to phages [10, 11]. In bacteria, CRISPR genes encode RNA arrays that can be processed to RNA fragments which bind Cas nuclease enzymes. These RNA fragments help guide the Cas nuclease to DNA sequences that are complementary to the Cas bound RNA. The CRISPR-Cas9 gene editing system employed in this protocol was developed by the Feng Zhang laboratory [12]. It is composed of the Cas9 nuclease and a chimeric RNA complex composed of three fused RNA sequences [13]: (1) the guide-RNA (gRNA) that is the user-determined sequence that targets a specific DNA sequence of interest, (2) a trans-activating crRNA (tracrRNA) required to form a complex with the Cas9 nuclease enzyme [14, 15], and (3) a CRISPR-RNA (crRNA) sequence that links the gRNA and tracrRNA [13, 16]. Through selection of a gRNA sequence specific to an early coding exon of a gene of interest, this system has been used to generate site-specific indels resulting in gene-specific knockouts.

The CRISPR-Cas9 plasmids developed by the Zhang lab contain the expression cassettes for both the Cas9 and CRISPR RNA complex [12]. A simple cloning step is required to insert a set of annealed short DNA oligonucleotides into this plasmid, encoding for the desired gRNA sequence, and specifying the gene targeting of the CRISPR-Cas9 enzyme [17]. Transfection and expression of this plasmid in cells initiates the gene editing experiment.

Although the system described above is efficient for many cell types, primary cells are less amenable to standard transfection protocols, especially without losing the primary characteristics of the cells. Primary human airway epithelial cells (AECs) are among such poorly transfectable cell types. To avoid this limitation the Zhang lab has cloned a modification of their CRISPR-Cas9 system into a lentiviral vector, allowing the delivery, stable integration, and expression of their system into a much wider range of primary cell types [18, 19]. This system also contains a puromycin resistance expression cassette to allow for selection of integrated cells, increasing the likelihood of editing and thus knockout within a transduced

cell population. This lentiviral-mediated CRISPR-Cas9 gene-editing system has recently been used for a genome-wide functional screen, as well as gene knockout studies in several human cell types with great success [18–22].

Since airway epithelial cells are amenable to lentiviral transduction, we tested the Zhang lentiviral CRISPR-Cas9 system as a potential method for gene editing of airway epithelial cells. Our application of this system to airway epithelial cells requires a brief delineation of the cell types and their characteristics that compose the airway epithelium. Airway epithelial cells include ciliated, secretory, and basal cell types. Basal airway epithelial cells serve as the stem cell of the airway. Consequently, they are the only AEC cell type capable of expansion in culture, are amenable to transduction, and can then be differentiated into a mucociliary epithelium using air–liquid interface (ALI) culture methods. Therefore, gene-editing of basal epithelial cells produces an expandable population of airway epithelial cells that can be cryopreserved for later expansion and generation of mucociliary cultures to be used in functional studies of a specific disease gene. We used the Zhang lab lentiviral CRISPR-Cas9 system, in combination with recently developed culture techniques [20, 23], to successfully knockout basal airway epithelial cells for the *MUC18* gene [20]. We then generated mucociliary epithelial cultures from these cells and characterized the functional significance of *MUC18* to various epithelial stimulations [20]. Since then, we have continued to improve the application of this lentiviral CRISPR-Cas9 gene-editing system to primary airway epithelial cells by refining our culture protocols in knocking out several other disease genes. Herein we describe these methods in detail in an effort to make this method accessible to all airway and lung disease researchers.

2 Materials

2.1 Cloning of a gRNA Sequence into the Lentiviral CRISPR Plasmid

1. Oligonucleotide sequences for gene target of interest (*see* Subheading 3.1.1).
2. Scrambled control oligonucleotide sequences (*see* Subheading 3.1.1).

2.1.1 Gene Target gRNA Selection and Oligonucleotide Design

2.1.2 gRNA Annealing and Golden Gate Cloning into LentiCRISPR Vector [24]

1. LentiCRISPR v2 backbone plasmid [19] (Addgene; Cambridge, MA; plasmid #52961).
2. Target DNA oligonucleotides.

3. Scrambled DNA oligonucleotides.
4. 10× T4 ligase buffer.
5. T4 polynucleotide kinase (PNK) enzyme (10,000 U/mL).
6. *BsmBI* endonuclease restriction enzyme (10,000 U/mL).
7. 2× rapid ligase buffer.
8. Bovine serum albumin (BSA) (10 mg/mL).
9. Quick T4 ligase.
10. Molecular grade water.
11. Sterile 0.2 mL PCR tubes
12. Thermal cycler.
13. Electrocompetent *E. coli* cells.
14. Luria–Bertani (LB) agar plates and broth supplemented with ampicillin (100 µg/mL).
15. Bacterial incubator at 37 °C.
16. Endotoxin-free Plasmid Miniprep Kit.
17. LKO.1 5' lentiCRISPR v2 plasmid sequencing primer.
 - (a) Sequence 5'-GACTATCATATGCTTACCGT-3'

2.2 Lentiviral Propagation

2.2.1 Generation of LentiCRISPR Lentivirus Using Lenti-X 293T Cells

1. Lenti-X 293T cells (Takara Bio USA).
2. Lenti-X 293T growth media: DMEM with 4.5 g/L glucose, L-glutamine, and sodium pyruvate, 10% heat-inactivated fetal bovine serum, 5.5 mL penicillin (10,000 IU)/streptomycin (10,000 µg/mL) solution, 5.5 mL L-glutamine (200 mM).
3. Lipofectamine 2000 transfection reagent.
4. OptiMEM reduced serum media.
5. pCMV-VSV-G pseudotyping plasmid [25] (Addgene; plasmid #8454).
6. psPAX2 lentiviral packaging plasmid (Addgene; plasmid #12260).
7. BEGM Bronchial Epithelial Cell Growth Media.
8. 100 mm tissue culture dishes.
9. Humidified tissue culture incubator at 37 °C with 5% CO₂.

2.2.2 Determination of Lentivirus Titer

1. Lenti-X qRT-PCR Titration Kit (Takara Bio, USA).
2. Sterile 96- or 384-well qPCR plates.
3. Thermal cycler.
4. Quantitative real-time PCR thermal cycler.

2.3 Transduction and Selection of Primary Basal Airway Epithelial Cells

2.3.1 Transduction of Basal Airway Epithelial Cells

1. Primary basal airway epithelial cells.
2. Rat tail collagen I (3 mg/mL).
3. Phosphate buffered saline (PBS) without calcium/magnesium.
4. BEGM Bronchial Epithelial Cell Growth Media.
5. Y-27632 dihydrochloride (10 mM stock solution).
6. HyClone HEPES (1 M) Buffer Solution.
7. Polybrene Infection/Transfection Reagent.
8. Lentivirus made with the lentiCRISPR-target-specific gRNA sequence of interest.
9. Lentivirus made with the lentiCRISPR-scramble gRNA sequence.
10. 100 mm tissue culture dishes.
11. Humidified tissue culture incubator at 37 °C with 5% CO₂.
12. Parafilm.

2.3.2 Selection and Harvest of Lentiviral-Transduced AECs

1. Puromycin dihydrochloride antibiotic stock solution (50 µg/mL in water; 0.2 µm filter-sterilized).
2. BEGM Bronchial Epithelial Cell Growth Media.
3. Y-27632 dihydrochloride (10 mM stock solution).
4. Phosphate buffered saline (PBS) without calcium/magnesium.
5. 0.25% trypsin–2.21 mM EDTA in HBSS.
6. Heat-inactivated fetal bovine serum (FBS).
7. Cryopreservation medium: 60% F-media (*see* Subheading 2.5.2), 30% heat-inactivated fetal bovine serum, 10% dimethyl sulfoxide, and Y-27632 dihydrochloride (10 µM).

2.4 Verification of CRISPR-Cas9 DNA Cutting by HRM Analysis

Use primer design software to generate PCR primer sequences to amplify a genomic DNA region covering the gene target gRNA cut site. Per our HRM reagents and instrument we typically attempt to design PCR products of 75–150 bp in length. Note that the expected cut site is 3–5 bp upstream of the PAM sequence (*see* **Notes 16 and 17**).

1. Genomic DNA Extraction Kit.
2. Screening primers.
3. MeltDoctor™ HRM Master Mix (Thermo Fisher Scientific).
4. QuantStudio™ 6 Flex Real-Time PCR System (Life Technologies) or equivalent.

2.5 Continued Selection and Harvest of Gene-Edited AECs

2.5.1 Preparation of Irradiated Fibroblast Feeder Layer

1. Puromycin resistant NIH/3T3 mouse embryonic fibroblasts (ATCC CRL-1658).
2. Fibroblast media: 500 mL DMEM containing glucose (4.5 g/L), L-glutamine, and sodium pyruvate, 50 mL heat-inactivated fetal bovine serum, 5.5 mL penicillin (10,000 IU)/streptomycin (10,000 µg/mL) solution, 5.5 mL L-glutamine (200 mM), 1 µg/mL puromycin dihydrochloride.
3. Phosphate buffered saline (PBS) without calcium/magnesium.
4. 0.25% trypsin–2.21 mM EDTA in HBSS.
5. Heat-inactivated fetal bovine serum (FBS).
6. Gamma Irradiator (Cesium-137 source).
7. 100 mm tissue culture treated dishes.

2.5.2 Continued Selection of Gene- Edited AECs

1. Cryopreserved CRISPR-transduced cells.
2. 37 °C water bath.
3. Complete DMEM Medium: 500 mL Dulbecco's Modified Eagle's Medium containing glucose (4.5 g/L), L-glutamine and without sodium pyruvate, 50 mL heat-inactivated fetal bovine serum, 5.5 mL penicillin (10,000 IU)/streptomycin (10,000 µg/mL) solution, 5.5 mL L-glutamine (200 mM).
4. Hydrocortisone/human Epidermal Growth Factor (HC/EGF) stock solution (1000×): Dissolve 2.5 µg of human Epidermal Growth Factor (Gibco) into 19 mL of DMEM containing glucose (4.5 g/L), L-glutamine and without sodium pyruvate. Prepare a 0.5 mg/mL solution of hydrocortisone in 100% molecular grade ethanol. Mix 1 mL of the hydrocortisone solution with 19 mL of the epidermal growth factor solution, filter-sterilize using a 0.2 µm filter, and store single use aliquots at –20 °C.
5. Cholera toxin stock solution: Prepare a 1 mg/mL solution of cholera toxin (Sigma-Aldrich) in molecular grade water, filter-sterilize using a 0.2 µm filter, and store aliquots at –20 °C.
6. Adenine stock solution: Prepare a 1.5 mg/mL solution in complete DMEM medium, and store 8 mL aliquots at –20 °C.
7. Insulin solution (Sigma-Aldrich #I9278).
8. Ham's F-12 Nutrient Mix medium.
9. F-media for basal cell growth and expansion: 365 mL complete DMEM medium, 125 mL Ham's F-12 Nutrient Mix, 8 mL adenine stock solution, 500 µL hydrocortisone/human epidermal growth factor (HC/EGF) stock solution, 500 µL insulin solution, 4.3 µL cholera toxin stock solution, filter-sterilize using a 500 mL 0.2 µm filter unit and store media at 4 °C for up to 1 month.

10. Puromycin dihydrochloride antibiotic stock solution: 50 µg/mL in water; 0.2 µm filter-sterilized.
11. Y-27632 dihydrochloride (10 mM stock solution).
12. Humidified tissue culture incubator at 37 °C with 5% CO₂.

2.5.3 Double Trypsinization Harvest of Cultured Airway Epithelial Cells

1. Phosphate buffered saline (PBS) without calcium/magnesium.
2. 0.25% trypsin–2.21 mM EDTA in HBSS.
3. Heat-inactivated fetal bovine serum (FBS).
4. Cryopreservation medium: 60% F-media (*see* Subheading 2.5.2), 30% heat-inactivated fetal bovine serum, 10% dimethyl sulfoxide, and Y-27632 dihydrochloride (10 µM).
5. Humidified tissue culture incubator at 37 °C with 5% CO₂.
6. Sterile 50 mL conical tubes.
7. Hemocytometer.
8. Trypan Blue solution: 0.4% Trypan Blue in PBS.

2.6 Final Harvest of Cells for Sequencing Analysis and Protein Knockout Validation

If sequence of the generated indels and indel frequency determination is desired, a massively parallel sequencing library can be generated for Ion Torrent sequencing using custom primers for two rounds of PCR. In the primary PCR reaction, adapters can be appended onto the gene-specific PCR primers (generated in Subheading 2.4) as described below.

2.6.1 Sequence Analysis of Indels by Ion Torrent Next-Generation Sequencing

1. Sequencing Library Generation—Primary PCR Primers
 - (a) Forward Primer: The universal adaptor sequence is appended onto the 5' end of the gene-specific forward primer as follows:
 - 5' CTGCTGTACGCAGCGT (Gene-specific Fwd primer sequence) 3'.
 - (b) Reverse primer: The trP1 sequencing adaptor for Ion Torrent sequencing is appended onto the 5' end of the gene-specific reverse primer as follows:
 - 5' CCTCTCTATGGGCAGTCGGTGAT (Gene-specific Rev. primer sequence) 3'.

Please note these primers will have to be generated for each specific cut site to be sequenced. However, these primers can be used for all cell sample PCR reactions for that cut site, since the barcoding will be introduced in the second PCR reaction.

2. Sequencing Library Generation—Secondary PCR Primers
 - (a) Forward primer: The Ion Torrent A-sequencing adapter followed by an Ion Torrent barcode, and then the universal adapter sequence:
 - 5' CCATCTCATCCCTGCGTGTCTCCGACTCAG-(Ion Torrent Barcode)-CTGCTGTACGCAGCGT 3'.

(b) Reverse primer: the trP1 sequencing adaptor sequence.

- 5' CCTCTCTATGGGCAGTCGGTGAT 3'.

Please note these secondary PCR primers are not cut site specific and thus can be used to amplify any PCR products generated from the primary PCR reaction, using a different barcode per cell sample.

3. Ion Torrent Personal Genome Machine (PGM) Sequencer.
4. PyroMark PCR Kit (Qiagen).
5. Purified genomic DNA.
6. Molecular grade water.
7. Sterile 0.2 mL PCR tubes.
8. Thermal cycler.
9. DNA electrophoresis tank.
10. 5× Tris–Borate–EDTA Buffer (TBE): 54 g Tris base, 27.5 g boric acid, and 20 mL of 0.5 M EDTA pH 8.0, bring to 1 L in deionized water and adjust solution to a final pH of 8.3.
11. Molecular biology agarose.
12. Phusion High-Fidelity DNA Polymerase (New England BioLabs).
13. DNA agarose gel extraction/purification kit.
14. Ion PGM Hi-Q View OT2 Kit (Thermo Fisher Scientific).
15. Ion PGM Hi-Q Sequencing Kit (Thermo Fisher Scientific).

3 Methods

3.1 Cloning of gRNA Sequence into the Lentiviral CRISPR Plasmid

3.1.1 Gene Target gRNA Selection and Oligonucleotide Design

To direct the CRISPR-Cas9 gene editing system to the gene targeted for knockout, a gRNA complementary to the DNA sequence for that gene must be designed. We suggest that the gRNA selection is performed with the Zhang lab CRISPR designer tool as recommended, to achieve efficient gene editing and knockout (*see Note 1*). The target region for gRNA design should be located in a coding exon of the studied gene. We typically submit the first three exons of a gene to the CRISPR design tool for selection of gRNA sequences. We note that gRNA sequences resulting from this system have the following characteristics:

1. The gRNA sequence is 20 bp in length.
2. A Protospacer Adjacent Motif (PAM) sequence of NGG is present at the 3'-end of the chosen target sequence [26].
3. A gRNA sequence with a “G” nucleotide located at the 5'-end is preferred (*see Note 2*).

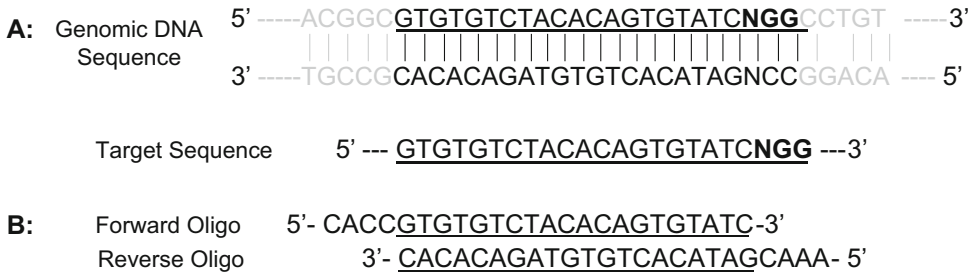


Fig. 1 Selection of the gRNA sequences and design of oligonucleotides based on these gRNAs for cloning into the lentiCRISPR plasmid. **(a)** The gene target gRNA sequence should be 20 bp in length directly upstream of a 3'-PAM NGG trinucleotide motif. If the gene target gRNA sequence does not begin with a “G” nucleotide, then it can be added to the oligonucleotide sequence to be synthesized (*see* **Notes 2** and **4**). **(b)** The oligonucleotides to be ordered for the gene target gRNA cloning. The NGG PAM sequence should not be included in the final oligonucleotide to be synthesized. For proper cloning into the lentiCRISPR backbone, the ligation adaptor sequence “CACC” should be added to the 5'-end of the forward oligonucleotide, while the ligation adaptor sequence “AAAC” should be added to the 5'-end of the reverse complement of the forward sequence which serves as the reverse oligonucleotide

Additionally, we randomly scramble the selected gene target gRNA sequence to generate a control gRNA sequence for our experiments (*see* **Note 3**). The gRNA sequences are incorporated into the lentiCRISPR vector through generation of gRNA sequence oligonucleotides complementary to one another. These oligonucleotides are annealed and then cloned into the lentiCRISPR vector as detailed below per the Zhang lab protocol.

Cloning of the annealed gRNA oligonucleotides requires attention to the following design details (*see* Fig. 1):

1. The oligonucleotide gRNA sequences should *NOT* contain the PAM (NGG) sequence.
2. The first gRNA oligonucleotide designed is the chosen target sequence (shown 5'→3' and designated “Forward Oligo”) and should have the nucleotide adaptor sequence CACC added to the 5'-end for ligation into the *BsmBI*-digested lentiCRISPR plasmid.
3. The second oligonucleotide should be the reverse complementary sequence to the 20 bp target sequence (shown 3'→5' and designated “Reverse Oligo”). The sequence AAAC should be added to the 5'-end of this sequence as the adaptor for ligation into the *BsmBI*-digested lentiCRISPR plasmid (*see* **Note 4**).

3.1.2 gRNA Annealing and Golden Gate Cloning in to LentiCRISPR Vector [24]

1. Assemble the reaction as indicated in Table 1 in 0.2 mL PCR tubes on ice to phosphorylate and anneal oligonucleotide sequences for each target or scrambled control of interest.
2. For 5' phosphorylation, incubate reaction mixture in a thermal cycler at 37 °C for 30 min followed by 95 °C for 5 min. Anneal

Table 1
Oligonucleotide phosphorylation and annealing reaction assembly

Reagent	Volume (μL)
Forward oligo (100 μM)	1
Reverse oligo (100 μM)	1
10 \times T4 ligase buffer	1
T4 PNK enzyme (10,000 U/mL)	0.5
Molecular grade H ₂ O	6.5
Total reaction volume	10

Table 2
Golden gate cloning reaction assembly

Reagent	Volume (μL)
2 \times rapid ligase buffer	12.5
BSA (10 mg/mL)	0.25
<i>BsmBI</i> endonuclease (10,000 U/mL)	1
Quick T4 ligase	0.125
Diluted annealed oligos (1:10)	1
lentiCRISPR v2 backbone vector (25 ng/ μL)	1
Molecular grade H ₂ O	9.125
Total reaction volume	25

phosphorylated oligonucleotides by ramping down the thermal cycler block temperature to 25 °C at 5 °C/min (*see Note 5*).

- Dilute each set of annealed oligonucleotides 1:10 by adding 90 μL of molecular grade water to each reaction tube.
- Assemble the Golden Gate cloning reaction on ice by mixing each set of annealed oligonucleotides with the following components in the order listed in Table 2.
- Incubate the Golden Gate cloning reaction in a thermal cycler using the following cycling conditions: 5 min at 37 °C (digestion) and 5 min at 20 °C (ligation), repeated for 15 cycles.
- Add 2 μL of the Golden Gate cloning reaction to electrocompetent *E. coli* cells and perform heat-shock transformation as per manufacturer's instructions (*see Note 6*).
- Plate transformed cultures onto LB/Ampicillin (100 $\mu\text{g}/\text{mL}$) agar plates for vector selection. Culture 2–3 individual colonies

from selection plates into LB/Ampicillin (100 µg/mL) broth for 16–20 h in a 37 °C shaking incubator and isolate the plasmid using an endotoxin-free plasmid miniprep kit (*see Note 7*).

- Verify insertion of the oligonucleotide guide sequence of each plasmid clone by sequencing using the LKO.1 5' sequencing primer.

3.2 Lentiviral Propagation (See Note 8)

Once the sequence verified gene specific and scrambled lenti-CRISPR vectors are constructed we generate lentivirus as detailed below. We generate lentivirus by using Lipofectamine 2000 and a modified version of the transfection protocol for Lenti-X 293T cells as described by Takara Bio USA.

3.2.1 Generation of LentiCRISPR Lentivirus Using Lenti-X 293T Cells

- Seed Lenti-X 293T cells at 5.5×10^4 cells/cm² in Lenti-X 293T growth media and incubate cells at 37 °C in 5% CO₂ for 24 h.
- Assemble transfection reaction A and reaction B in separate tubes as shown in Table 3.
- Mix transfection reaction A and B together and incubate for 20 min at room temperature. Add the total volume dropwise to the existing growth media in Lenti-X 293T cell culture dishes, gently rotating the dish forward and back, to ensure even mixing. Incubate cells in a humidified tissue culture incubator at 37 °C with 5% CO₂ (*see Note 9*).
- After 24 h, replace Lenti-X 293T growth media from cells with complete BEGM growth media and return to tissue culture incubator.

Table 3
Lenti-X 293T transfection reaction assembly

Reagent	Amount
Reaction A	
Verified lentiCRISPR plasmid	7000 ng
psPAX2 packaging plasmid	9000 ng
pCMV-VSV-G pseudotyping plasmid	900 ng
OptiMEM reduced serum media	To 850 µL
Reaction B	
Lipofectamine 2000 transfection reagent	25 µL
OptiMEM reduced serum media	To 850 µL

5. Incubate for an additional 48 h and harvest viral-containing cell culture media. Remove any intact cells and cell debris by centrifugation at $225 \times g$ for 5 min at 4 °C.
6. Aliquot supernatant containing lentiviral-media and store at -80 °C until use.

3.2.2 Determination of Lentivirus Titer

1. Thaw an aliquot of lentiviral-media on ice.
2. Extract viral RNA, set up the lentiviral titration reaction, and calculate the lentiviral titer (lentiviral copies/mL) using the Lenti-X™ qRT-PCR Titration Kit as per manufacturer's instructions (*see* **Note 10**).

3.3 Transduction and Selection of Primary Basal Airway Epithelial Cells

3.3.1 Transduction of Basal Airway Epithelial Cells

Once high titer CRISPR-Cas9 lentivirus guided by the gene target gRNA and the scrambled control gRNA are generated, transduction of basal airway epithelial cells is performed as detailed below. We suggest that passage 1 or passage 2 basal airway epithelial cells be used for viral transduction. We also suggest the use of donor cells that have been previously verified to robustly proliferate in submerged BEGM culture and differentiate well using air-liquid interface culture.

1. First, prepare collagen coated 100 mm tissue culture dishes for seeding of basal airway epithelial cells to be transduced by each gene target gRNA and scramble control gRNA guided CRISPR-Cas9 lentivirus. Mix 42.5 μ L rat tail collagen I with 5 mL of PBS by vortexing for 10 s, and then add to the culture vessel. After a 45 min incubation at room temperature, gently wash the dish 2 times with PBS and allow the collagen coating to air-dry prior to plating cells.
2. Seed basal airway epithelial cells on collagen coated plates at a density of 5.5×10^3 cells/cm² in complete BEGM growth media supplemented with Y-27632 (10 μ M), and incubate for 48 h until cells are 30–40% confluent (*see* **Notes 11** and **12**). Each experiment requires one plate of seeded basal airway epithelial cells for transduction with the scrambled control gRNA guided CRISPR-Cas9 lentivirus and one plate per gene target gRNA guided CRISPR-Cas9 lentivirus being tested.
3. For lentiviral transduction, combine 15 μ L Polybrene, 200 μ L 1 M HEPES, 3×10^8 copies of lentivirus (from lentiviral titration in Subheading 3.2.2), and add complete BEGM growth media for a total transduction reaction volume of 10 mL.
4. Remove cell culture media from plated basal airway epithelial cells and carefully add the lentiviral transduction mix to the cell culture dish of epithelial cells (*see* **Note 9**).

5. Centrifuge tissue culture dishes at $920 \times g$ for 1 h at room temperature to increase the transduction efficiency (*see Note 13*).
6. Remove and discard the transduction media and replace with BEGM growth media supplemented with Y-27632 (10 μ M) (*see Note 9*).
7. Return the lentiviral-transduced basal airway epithelial cells to a humidified tissue culture incubator at 37 °C with 5% CO₂.

3.3.2 Selection and Harvest of Lentiviral-Transduced AECs

At this point in the protocol we harvest the transduced cells to evaluate cutting efficiency mediated by the tested gRNAs. The harvested cells are partitioned for both genomic DNA extraction (used in HRM assays) and for cryopreservation.

1. At 24 h post-lentiviral transduction, add puromycin (final concentration of 1 μ g/mL) to the existing growth media to initiate selection of lentiviral-integrated cells (*see Note 14*). Change BEGM growth media supplemented with Y-27632 (10 μ M) and puromycin (1 μ g/mL) every other day until cells reach 90% confluence.
2. To harvest transduced epithelial cells, remove cell culture media and wash adhered cells with PBS; discard PBS wash.
3. Add 5 mL of prewarmed 37 °C trypsin to the cell monolayer, and incubate in a tissue culture incubator for 5 min until cells detach from the dish.
4. Collect cells and neutralize trypsin by the addition of 1 mL of FBS. Wash dish with 10 mL of PBS to ensure complete harvest of all the cells, and pool with the previously collected sample.
5. Pellet at $225 \times g$ for 5 min at 4 °C; resuspend cell pellet in 1 mL of PBS and determine cell count.
6. From harvested cells, process each harvested cell population as follows:
 - (a) Add desired number of cells to genomic DNA lysis buffer. Process immediately or store at -80 °C until extraction of DNA. This DNA will be used as the template for high resolution melt curve (HRM) analysis as described in Subheading 3.4 (*see Note 15*).
 - (b) Prepare cells at a concentration between 5×10^5 and 1×10^6 cells per vial in cryopreservation media for long-term storage in liquid nitrogen and later continued selection as described in Subheading 3.5.

The simplest way to evaluate whether the tested gRNAs mediate efficient DNA cutting is to perform High Resolution Melt curve (HRM) analysis of a PCR product generated (from the genomic DNA of each CRISPR edited cell sample) across the gRNA cut site.

3.4 Verification of CRISPR-Cas9 DNA Cutting by HRM Analysis

If the particular gRNA guided Cas9 enzyme is cutting the DNA at the designed site, indels will be created and thus multiple species of PCR products will be generated. In contrast, the PCR product generated from the scrambled control gRNA cells will contain only one species of PCR product (the unedited species). The melting profile of the PCR products generated from the gene target gRNA and scramble control gRNA cells will then differ by HRM analysis.

1. Extract DNA from DNA lysates using a genomic DNA extraction kit as per manufacturer's instructions.
2. Assemble the HRM qPCR reaction on ice as described in Table 4 (*see* Notes 16 and 17). PCR primers should be designed as detailed in Subheading 2.4.
3. The HRM analysis program should be run on a QuantStudio™ 6 Flex Real-Time PCR System or equivalent using the cycling conditions listed in Table 5.
4. HRM analysis interpretation: By comparing the melt curves of PCR products amplified over the cut site of DNA from the

Table 4
HRM analysis reaction assembly

Reagent	Amount
MeltDoctor™ HRM master mix (2×)	2.5 μL
Forward screening primer (5 μM)	0.3 μL
Reverse screening primer (5 μM)	0.3 μL
Genomic DNA	5 ng
Molecular grade H ₂ O	To 5 μL total volume

Table 5
PCR and HRM melt curve analysis cycling conditions

Stage	Step	Temp (°C)	Time	Ramp rate (°C/s)
Holding	Enzyme activation	95	10 min	1.6
Cycling (40 cycles)	Denature	95	15 s	1.6
	Anneal/extend	60	1 min	1.6
High resolution melt curve	Denature	95	10 s	1.6
	Anneal	60	1 min	1.6
	High resolution melt	95	15 s	0.025
	Anneal	60	15 s	1.6

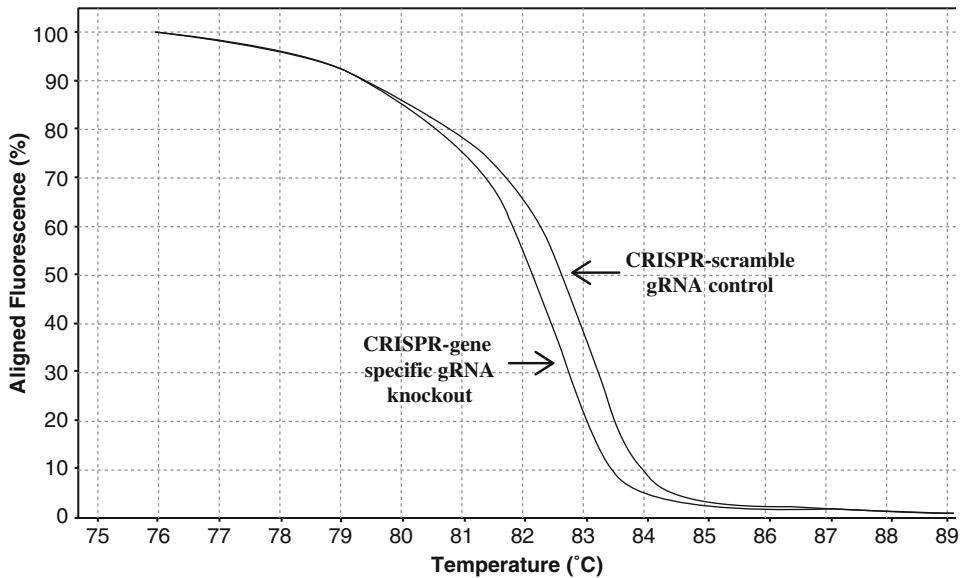


Fig. 2 High resolution melt curve analysis of PCR products amplified over the cut site, generated from DNA isolated from scrambled control gRNA and gene target gRNA cells. The shift left of the gene target gRNA cells indicates the presence of multiple PCR products resultant from multiple indels at the cut site, that in sum have a different melt profile from the scrambled control gRNA cells

scrambled gRNA control and gene target gRNA cell populations, it can be determined whether DNA cutting has occurred in the targeted gene cut site (*see* Fig. 2). A leftward shift in the melt curve of the gene target gRNA cells compared to the scrambled control gRNA cells indicates the presence of indels within the targeted DNA region. The size of the shift is indicative of the proportion of cells that underwent DSB and NHEJ, resulting in the insertion/deletion of nucleotides at the cut site.

3.5 Continued Selection and Harvest of Gene-Edited AECs

Following HRM analysis to verify the DNA cutting efficiency of each gRNA used to target the gene of interest, it is suggested to select the most efficient gRNA to proceed with downstream experiments. The gene target gRNA that results in the most significant shift from the scramble control gRNA in HRM curve analysis is likely to contain the most cells with bi-allelic gene knockout. Therefore, we select these gRNA treated cells for further selection. We have observed that continued selection of this population of cells using a modified Schlegel culture method for two additional passages generates the maximum level of gene knockout for this bulk-selected population [20, 23]. The method described here involves growth and selection of the edited basal airway epithelial cells using a modified Schlegel culture method [20, 23, 27, 28]. Briefly, this method involves epithelial cell culture on an irradiated fibroblast feeder layer using specialized growth media supplemented with Y-27632.

3.5.1 Preparation of Irradiated Fibroblast Feeder Layer

Since the edited basal airway epithelial cells will continue to be cultured with puromycin for continued selection, puromycin resistant fibroblasts must be used to generate the feeder layer for culture. The generation of puromycin resistant fibroblast feeder cells can be done by transducing NIH/3T3 mouse embryonic fibroblasts with an empty lentiCRISPR vector backbone as described previously [20]. Irradiation and seeding of puromycin resistant fibroblasts can be completed as described below.

1. Culture puromycin resistant NIH/3T3 mouse embryonic fibroblasts in prewarmed Fibroblast media until cells are ~80% confluent.
2. To harvest fibroblasts, remove cell culture media and wash adhered cells with PBS; discard PBS wash.
3. Add 5 mL of prewarmed 37 °C trypsin to the cell monolayer, and incubate in a tissue culture incubator for 3–5 min or until cells detach from the dish.
4. Collect cells and neutralize trypsin by the addition of 1 mL of FBS. Wash dish with 10 mL of PBS to ensure complete harvest of all the cells, and pool with the previously collected sample.
5. Pellet at $225 \times g$ for 5 min at 4 °C; resuspend cell pellet in 1 mL of PBS and determine cell count.
6. Subculture between 1.3×10^3 and 6.6×10^3 cells/cm² in a tissue culture dish in Fibroblast media for continued culture and expansion.
7. Irradiate the remaining cells by exposure to 5000 rads of gamma radiation using a Cesium-137 radiation source.
8. Seed irradiated puromycin resistant fibroblasts at 2.7×10^4 cells/cm² into a 100 mm tissue culture dish in Fibroblast media, and incubate in a tissue culture incubator at 37 °C with 5% CO₂ for 24 h.
 - (a) Irradiated cell monolayers should be ~90% confluent at 24 h post-seeding and should be used within 24–72 h after plating for the optimal seeding and expansion of seeded basal airway epithelial cells.

3.5.2 Continued Selection of Gene-Edited AECs

1. Quick thaw the gene-edited AECs removed from liquid nitrogen storage in a 37 °C water bath just until all ice crystals have melted.
2. Wash cells by adding total volume of thawed cells to 9 mL of prewarmed 37 °C F-media. Pellet cells at $225 \times g$ for 5 min at 4 °C; discard supernatant, taking care to not dislodge cell pellet.

3. Resuspend pellet in 10 mL of prewarmed 37 °C F-media supplemented Y-27632 (10 μM) and puromycin (1 μg/mL) for culture and AEC selection.
4. Using a 100 mm dish seeded with irradiated puromycin resistant NIH/3T3 mouse embryonic fibroblasts (<3 days since seeding), remove culture media. Gently add the basal airway epithelial cell suspension to the side of the dish and return seeded cells to a tissue culture incubator at 37 °C with 5% CO₂ (*see Note 9*).
5. Culture cells for up to, but no more than, 12 days checking for the development of epithelial cell colonies and conducting media changes every 48 h with F-media supplemented with Y-27632 (10 μM) and puromycin (1 μg/mL) until cell harvest (*see Note 18*).

3.5.3 Double Trypsinization Harvest of Cultured Airway Epithelial Cells

1. Aspirate culture media and wash cell monolayer with 10 mL of PBS; discard wash.
2. To first remove the irradiated fibroblast feeder layer, add 5 mL of prewarmed 37 °C trypsin to the dish. Incubate dish in tissue culture incubator for *exactly* 1 min (*see Note 19*).
3. To wash away fibroblasts while leaving the epithelial cells attached, immediately remove trypsin and wash epithelial monolayer with 10 mL of room temperature PBS and discard wash. Repeat wash for a second time, discard PBS wash, and microscopically observe culture to ensure that >90% of the irradiated fibroblast cells have been removed.
4. To harvest basal epithelial cell colonies, add 5 mL of prewarmed 37 °C trypsin to the dish and incubate in a tissue culture incubator at 37 °C with 5% CO₂ for 5 min (*see Note 20*).
5. To collect cells, add 5 mL of PBS to the existing trypsin in the dish and wash the cells from the dish surface by resuspending the 10 mL trypsin/cell/PBS mixture 2–5 times. Add the cell suspension to a 50 mL conical tube.
6. Neutralize trypsin by immediately adding 1 mL of heat-inactivated fetal bovine serum; mix well.
7. Wash the dish with 10 mL of PBS to ensure all epithelial cells have been dislodged and collected; pool wash with neutralized cell suspension.
8. Pellet cells at 225 × *g* for 5 min at 4 °C; discard supernatant, taking care to not dislodge cell pellet.
9. Suspend pellet in 1 mL of F-media supplemented Y-27632 (10 μM). Prepare a 1:10 dilution of the cell suspension in Trypan Blue solution and determine cell counts on a hemocytometer.

10. From expanded and harvested cells, process each cell line as follows:
 - (a) Subculture 5×10^5 cells one more passage by following Subheading 3.5.2, steps 3–5.
 - (b) Prepare cell lines at a concentration between 5×10^5 and 1×10^6 cells per vial in cryopreservation media for long-term storage in liquid nitrogen.

3.6 Final Harvest of Cells for Sequencing Analysis and Protein Knockout Validation

At this stage in the protocol, following two rounds of selection on an irradiated fibroblast feeder layer, we believe the transduced and selected basal airway epithelial cells are both fully selected and have had ample time to allow CRISPR-Cas9 cutting and subsequent indel formation to occur. DNA harvested at this stage from the scrambled control gRNA and gene target gRNA-transduced basal airway epithelial cells may be analyzed by next-generation sequencing to determine the frequency of indels at the cut site, as described below in Subheading 3.6.1.

Depending on the objectives of the experiment, basal cells harvested here can also be immediately analyzed or further expanded for additional experiments. If the gene of interest is expressed in basal airway epithelial cells, these expanded cells may be used for validation of gene knockout by a variety of methods including Western Blot analysis, immunofluorescence staining, and flow cytometry. Moreover, these cells can be used in experiments to determine the function of the gene that has been disrupted. To determine the function of this gene in a mucociliary epithelium or if the studied gene is only expressed in the mucociliary epithelium, air–liquid interface cultures will need to be generated from these gene-edited basal cells. The ALI culture protocol we suggest is the one we describe in Reynolds et al. [23] with the following adaptations. First, due to the high level of selection and passage with which the gene-edited cells have undergone up to this point, we suggest seeding selected basal cells at a higher density of 3.0×10^5 cells/cm² onto transwell inserts. Secondly, once these seeded cells have expanded and reached confluence on the transwell inserts (using the Reynolds protocol), we recommend switching to PneumaCult-ALI Medium (StemCell Technologies) for air–liquid interface differentiation of the cultures. We suggest allowing PneumaCult-ALI differentiation to proceed for 21 days to ensure the cells are well differentiated before starting experimentation.

3.6.1 Sequence Analysis of Indels by Ion Torrent Next-Generation Sequencing

Although HRM analysis provides evidence that the genomic DNA isolated from the treated and selected cells is cut and contains indels at the designed cut site, it does not reveal the percentage of DNA alleles and thus cells that have indels or the particular indel sequences. We determine this by performing massively parallel sequencing of the same PCR product amplified over the cut site

Table 6
Primary PCR reaction assembly for sequence analysis

Reagent	Volume (μL)
PyroMark master mix ($2\times$)	5
Coral load ($10\times$)	1
Forward primer ($12.5\ \mu\text{M}$)	0.16
Reverse primer ($12.5\ \mu\text{M}$)	0.16
Genomic DNA ($10\ \text{ng}/\mu\text{L}$)	2
Molecular grade water	1.68
Total reaction volume	10

that is used in the HRM analysis. Specifically, we use Ion Torrent sequencing on the Personal Genome Machine (PGM). The PGM sequence generation scale, price, and time make this instrument perfect for indel sequence analysis. We have developed a custom protocol to generate barcoded sequencing libraries for each edited cell population generated. These barcoded libraries can be combined and run on PGM sequencing chips. Our library generation protocol involves two rounds of PCR amplification amplifying the cut site with specially designed primers to generate barcoded library molecules as detailed below.

1. Assemble the primary PCR reaction on ice for each DNA sample (e.g., DNA isolated from each scrambled control gRNA and each gene target gRNA cell sample) as indicated in Table 6. Primers should be designed as detailed in Subheading 2.6.1. Please note that only one set of PCR primers is needed per cut site amplified, regardless of the number of samples to combine on a sequencing run, due to the universal adapter sequence contained in the forward primer that allows barcoding in the second PCR step.
2. For PCR amplification, incubate the reaction in a thermal cycler using the following cycling conditions: 15 min at $95\ ^\circ\text{C}$, 35 cycles of $94\ ^\circ\text{C}$ for 30 s, $60\ ^\circ\text{C}$ for 30 s, $72\ ^\circ\text{C}$ for 30 s, followed by 10 min at $72\ ^\circ\text{C}$, and hold at $4\ ^\circ\text{C}$.
3. To confirm the size and specificity of the primary PCR amplicon, prepare a 1.8% agarose gel in $1\times$ TBE buffer, and load and run $2\ \mu\text{L}$ of each PCR product by gel electrophoresis (*see Note 21*).
4. To add the A-sequencing adaptor and to add individual sequencing barcodes to each sample, assemble the second PCR reaction on ice using the primary PCR product from each DNA sample in the order listed as indicated in Table 7.

Table 7
Secondary PCR reaction assembly for sequence analysis

Reagent	Volume (μL)
5 \times High Fidelity buffer	2
dNTPs (10 mM)	0.2
Forward primer (A-seq/barcode X) (5 μM)	0.5
Reverse primer (trP1 adaptor) (5 μM)	0.5
Phusion polymerase	0.1
PCR product (diluted 1:10 in water)	1
Water	5.7
Total reaction volume	10

A single reverse primer can be used for all cell samples, for each cut site examined. However, a different forward PCR primer is needed for each cell sample, each with a different barcode, if you intend to combine all libraries in a single sequencing run.

5. For PCR amplification of the second PCR reaction, incubate the reaction in a thermal cycler using the following cycling conditions: 2 min at 98 °C, 35 cycles of 98 °C for 10 s, 60 °C for 20 s, 72 °C 30 s, followed by 5 min at 72 °C, and hold at 4 °C.
6. To verify amplification and desired amplicon size of the second PCR product, prepare a 2% agarose gel in 1 \times TBE buffer, and load and run 5 μL of each PCR product by gel electrophoresis.
7. Extract and gel purify the correct size band from each lane and quantify each sample.
8. Pool equal molar amounts of each purified barcoded sample (ensuring that each sample has a different barcode sequence) for the library for PGM sequencing.
9. Prepare sequencing templating reaction as per manufacturer's protocol (Thermo Fisher Scientific manual MAN0014579) and perform setup and sequencing of the sample library pool on the Ion Torrent Personal Genome Machine (PGM) as per manufacturer's instructions (Thermo Fisher Scientific manual MAN0009816)
10. Bioinformatic analysis of sequencing data for the percentage of gene-edited reads and sequence of indels can be completed as described by Chu et al. [20].

4 Notes

1. The Zhang lab CRISPR gRNA sequence design tool is offered at (<http://crispr.mit.edu/>). The design tool requires the nucleotide sequence of the gene target site (up to 500 bp). The program identifies gRNA sequences, provides quality scores for each identified gRNA, and lists each candidate's potential genome-wide off-targets. It is suggested to select and clone at least three different gRNA sequences for the target gene, as some gRNAs result in better editing efficiencies than others.
2. The U6 promoter responsible for the transcription of the CRISPR RNA cassette is more efficient with a 5' "G" nucleotide at the transcription start site. If the selected 20 bp gRNA sequence does not have a "G" nucleotide at the 5'-end, an additional "G" nucleotide should be added to the 5'-end of the gRNA sequence resulting in a 21 bp gRNA sequence.
3. We recommend that the selected gene target gRNA sequence be scrambled and this scrambled gRNA tested in parallel to the target gene gRNAs. The experimental results of the gene KO cells can then be compared to scrambled gRNA cells, which serve as an appropriate control for the full experimental cycle used to generate the gene KO cells. The oligonucleotide design required for cloning this scrambled control gRNA sequence into the lentiCRISPR vector should be followed as described for the gene target gRNA sequence in Subheading 3.1.2.
4. If an additional "G" nucleotide is added to the target sequence as instructed in *see* **Note 2**, ensure that a "C" nucleotide is added to the 3'-end of the reverse oligonucleotide sequence to maintain complementarity of the two oligonucleotide sequences.
5. If ramp down temperature option is not available on the thermal cycler, similar annealing results will occur if the sample is placed on the bench and allowed to slowly cool to room temperature.
6. It is not necessary to include a negative control in the Golden Gate assembly reaction and transformation, as the empty lentiCRISPR backbone will religate with itself and result in empty vector-containing bacterial colonies under selection.
7. Alternative kits for plasmid isolation may be used, however endotoxin-free plasmid preparation is recommended as downstream use of isolated constructs will be used for mammalian cell transfection experiments.
8. The methods and propagation of virus described here use a 2nd generation lentiviral vector system that depends on a three-plasmid approach to produce a replication incompetent

lentivirus, barring recombination events. We use BSL-2 precautions in all steps involving virus generation and use of virus. Users of this protocol should first develop an appropriate biosafety plan and obtain approval from their institutional biosafety and other relevant committees. Part of these biosafety precautions should be the use of appropriate personal protective equipment (PPE) at all times.

9. Lenti-X 293T cells, airway epithelial cells, and irradiated NIH/3T3 mouse embryonic fibroblasts are easily detached from the bottom of tissue culture vessels during media changes and the addition of reagents. Take care to add reagents carefully to the edge of the dish, as disturbing the cell monolayer will adversely affect the titer of virus produced by the cells, the outcome of lentiviral transduction of epithelial cells, and the growth of sub-cultured airway epithelial cells, respectively.
10. Lentiviral titer qPCR assays can be completed in either 96-well or 384-well plates, as long as the reaction volumes and viral titer calculations are properly scaled using the calculation formula provided with the Lenti-X™ qRT-PCR Titration Kit.
11. The seeding density and incubation times for lentiviral transduction have been validated for nasal-, tracheal-, and bronchial-derived basal airway epithelial cells ([20], unpublished data). The protocol for cells derived from other sources may need to be validated for optimal transduction and gene-editing results.
12. Cells should not be grown past 30–40% confluency during this step. Overconfluent cultures will affect the lentivirus-to-cell ratio, negatively affecting the transduction efficiency and antibiotic selection of lentiviral integrated cells.
13. For biosafety purposes, lentiviral-containing tissue culture dishes must be properly handled and secured to prevent contamination of equipment. Prior to centrifugation, carefully wrap each tissue culture dish in Parafilm to seal the edges of each dish. Ensure that the dishes are placed in the appropriate vessel or rotor bucket and properly secured before starting the centrifuge.
14. The use of puromycin antibiotic at a concentration of 1 µg/mL was empirically determined and optimized for selection of nasal-, tracheal-, and bronchial-derived epithelial cells ([20], unpublished data). The use of other selection antibiotics or cells derived from alternative sources will require optimization of the selection protocol and concentration of the selection agent prior to use.
15. A minimum of 5×10^4 cells should be collected for genomic DNA extraction, although the number of cells can vary based on the extraction kit and type of cells used. Cell stocks should

be cryopreserved in liquid nitrogen according to standard laboratory protocols for later expansion.

16. High resolution melt curve analysis compares the melt curves of PCR products generated over the cut site between the scrambled control gRNA cells versus the gene target gRNA cells. Reactions must be set up using: (1) DNA from the scrambled control gRNA cells, and (2) DNA from the gene target gRNA cells, to interpret HRM analysis as described in this protocol. Samples should be run in duplicate.
17. High Resolution Melt analysis primers must be sequence specific. If melt curve analysis identifies multiple amplicon products, primer sets should be redesigned to yield a single amplicon during PCR amplification for accurate analysis.
18. The confluence of airway epithelial cells on the irradiated fibroblast feeder layer will vary based on a number of factors including: (1) the number of cells seeded, (2) the level of selection the initial transduced cell population underwent, (3) and the efficiency of lentiviral integration. It is imperative the cells should be allowed to grow for no more than 12 days as the irradiated fibroblast feeder layer will begin to deteriorate and affect the growth of the basal airway epithelial cell colonies. Likewise, epithelial colonies should not be allowed to grow to more than ~80% total confluence on the plate and special care should be taken to not let each individual colony get too large as to risk undue stress and replication cycles on the continuously expanding cell population.
19. It is important that this first trypsinization step be kept to *exactly* 1 min of duration in the 37 °C tissue culture incubator. Upon microscopic inspection, this 1 min step will allow for the release of the loosely bound irradiated fibroblasts, while maintaining the adherence of the basal airway epithelial cell colonies. Incubation longer than 1 min will have an adverse effect on the harvested number of basal epithelial cells as they will begin to lose their adherence and be removed with subsequent wash steps.
20. Epithelial cell trypsinization time can vary based on the cell number, density, and length of previous wash steps. This second trypsinization step should not be allowed to progress more than 7 min as longer incubation begins to have a negative effect on the viability and subsequent culture of the basal airway epithelial cells.
21. For downstream sequencing analysis, it is important that a single correctly sized PCR amplicon is produced from this primary PCR reaction. Moving forward into the second PCR with a multibanded product may result in poor PCR and sequencing results. Larger bands may be indicative of

nonspecific amplification, in which case different amplification primers may need to be designed. A bright <100 bp band is indicative of primer dimers that were formed during the reaction. In the case of larger (nonspecific) or smaller (primer dimer) products, the correct sized DNA product must be gel excised and purified prior to setup of the second PCR reaction.

References

1. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391(6669):806–811. <https://doi.org/10.1038/35888>
2. Ramachandran S, Krishnamurthy S, Jacobi AM, Wohlford-Lenane C, Behlke MA, Davidson BL, PB MC Jr (2013) Efficient delivery of RNA interference oligonucleotides to polarized airway epithelia in vitro. *Am J Physiol Lung Cell Mol Physiol* 305(1):L23–L32. <https://doi.org/10.1152/ajplung.00426.2012>
3. Hammond SM (2005) Dicing and slicing: the core machinery of the RNA interference pathway. *FEBS Lett* 579(26):5822–5829. <https://doi.org/10.1016/j.febslet.2005.08.079>
4. Tomari Y, Zamore PD (2005) Perspective: machines for RNAi. *Genes Dev* 19(5):517–529. <https://doi.org/10.1101/gad.1284105>
5. Mocellin S, Provenzano M (2004) RNA interference: learning gene knock-down from cell physiology. *J Transl Med* 2(1):39. <https://doi.org/10.1186/1479-5876-2-39>
6. Boch J, Scholze H, Schornack S, Landgraf A, Hahn S, Kay S, Lahaye T, Nickstadt A, Bonas U (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* 326(5959):1509–1512. <https://doi.org/10.1126/science.1178811>
7. Hurt JA, Thibodeau SA, Hirsh AS, Pabo CO, Joung JK (2003) Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection. *Proc Natl Acad Sci U S A* 100(21):12271–12276. <https://doi.org/10.1073/pnas.2135381100>
8. Wright DA, Thibodeau-Beganny S, Sander JD, Winfrey RJ, Hirsh AS, Eichtinger M, Fu F, Porteus MH, Dobbs D, Voytas DF, Joung JK (2006) Standardized reagents and protocols for engineering zinc finger nucleases by modular assembly. *Nat Protoc* 1(3):1637–1652. <https://doi.org/10.1038/nprot.2006.259>
9. Zhang F, Cong L, Lodato S, Kosuri S, Church GM, Arlotta P (2011) Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat Biotechnol* 29(2):149–153. <https://doi.org/10.1038/nbt.1775>
10. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* 60(2):174–182. <https://doi.org/10.1007/s00239-004-0046-3>
11. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315(5819):1709–1712. <https://doi.org/10.1126/science.1138140>
12. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* 339(6121):819–823. <https://doi.org/10.1126/science.1231143>
13. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337(6096):816–821. <https://doi.org/10.1126/science.1225829>
14. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471(7340):602–607. <https://doi.org/10.1038/nature09886>
15. Karvelis T, Gasiunas G, Miksys A, Barrangou R, Horvath P, Siksnys V (2013) crRNA and tracrRNA guide Cas9-mediated DNA interference in *Streptococcus Thermophilus*. *RNA Biol* 10(5):841–851. <https://doi.org/10.4161/rna.24203>
16. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuys RJ, Snijders AP, Dickman MJ,

- Makarova KS, Koonin EV, van der Oost J (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* 321 (5891):960–964. <https://doi.org/10.1126/science.1159689>
17. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F (2013) Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* 8 (11):2281–2308. <https://doi.org/10.1038/nprot.2013.143>
 18. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen T, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* 343(6166):84–87. <https://doi.org/10.1126/science.1247005>
 19. Sanjana NE, Shalem O, Zhang F (2014) Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods* 11 (8):783–784. <https://doi.org/10.1038/nmeth.3047>
 20. Chu HW, Rios C, Huang C, Wesolowska-Andersen A, Burchard EG, O'Connor BP, Fingerlin TE, Nichols D, Reynolds SD, Seibold MA (2015) CRISPR-Cas9-mediated gene knockout in primary human airway epithelial cells reveals a proinflammatory role for MUC18. *Gene Ther* 22(10):822–829. <https://doi.org/10.1038/gt.2015.53>
 21. Bellec J, Bacchetta M, Losa D, Anegón I, Chanson M, Nguyen TH (2015) CFTR inactivation by lentiviral vector-mediated RNA interference and CRISPR-Cas9 genome editing in human airway epithelial cells. *Curr Gene Ther* 15(5):447–459
 22. Firth AL, Menon T, Parker GS, Qualls SJ, Lewis BM, Ke E, Dargitz CT, Wright R, Khanna A, Gage FH, Verma IM (2015) Functional gene correction for cystic fibrosis in lung epithelial cells generated from patient iPSCs. *Cell Rep* 12(9):1385–1390. <https://doi.org/10.1016/j.celrep.2015.07.062>
 23. Reynolds SD, Rios C, Wesolowska-Andersen A, Zhuang Y, Pinter M, Happoldt C, Hill CL, Lallier SW, Cosgrove GP, Solomon GM, Nichols DP, Seibold MA (2016) Airway progenitor clone formation is enhanced by Y-27632-dependent changes in the Transcriptome. *Am J Respir Cell Mol Biol* 55 (3):323–336. <https://doi.org/10.1165/rcmb.2015-0274MA>
 24. SAM target sgRNA cloning protocol (2014) <http://sam.genome-engineering.org/protocols/>
 25. Stewart SA, Dykxhoorn DM, Palliser D, Mizuno H, EY Y, An DS, Sabatini DM, Chen IS, Hahn WC, Sharp PA, Weinberg RA, Novina CD (2003) Lentivirus-delivered stable gene silencing by RNAi in primary cells. *RNA* 9 (4):493–501
 26. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* 151 (Pt 8):2551–2561. <https://doi.org/10.1099/mic.0.28048-0>
 27. Suprynowicz FA, Upadhyay G, Krawczyk E, Kramer SC, Hebert JD, Liu X, Yuan H, Cheluvvaraju C, Clapp PW, Boucher RC Jr, Kamonjoh CM, Randell SH, Schlegel R (2012) Conditionally reprogrammed cells represent a stem-like state of adult epithelial cells. *Proc Natl Acad Sci U S A* 109 (49):20035–20040. <https://doi.org/10.1073/pnas.1213241109>
 28. Liu X, Ory V, Chapman S, Yuan H, Albanese C, Kallakury B, Timofeeva OA, Nealon C, Dakic A, Simic V, Haddad BR, Rhim JS, Dritschilo A, Riegel A, McBride A, Schlegel R (2012) ROCK inhibitor and feeder cells induce the conditional reprogramming of epithelial cells. *Am J Pathol* 180(2):599–607. <https://doi.org/10.1016/j.ajpath.2011.10.036>

Chapter 16

RNA Interference to Knock Down Gene Expression

Haiyong Han

Abstract

RNA interference (RNAi) is a biological process by which double-stranded RNA (dsRNA) induces sequence-specific gene silencing by targeting mRNA for degradation. As a tool for knocking down the expression of individual genes posttranscriptionally, RNAi has been widely used to study the cellular function of genes. In this chapter, I describe procedures for using gene-specific, synthetic, short interfering RNA (siRNA) to induce gene silencing in mammalian cells. Protocols for using lipid-based transfection reagents and electroporation techniques are provided. Potential challenges and problems associated with the siRNA technology are also discussed.

Key words RNA interference, RNAi, siRNA, Gene silencing, Transfection, Electroporation

1 Introduction

Specific inhibition or knockdown of gene expression in cultured cells has been widely used to study the effects of loss-of-function mutation in individual genes. Gene-specific degradation of mRNA is one way to silence individual gene expression posttranscriptionally. One of the most widely used technologies for induction of such gene-specific RNA degradation is the use of RNA interference (RNAi) technology. RNAi was first discovered in the nematode *C. elegans* as a response to small double-stranded RNA (dsRNA), which resulted in sequence-specific gene silencing [1].

RNAi is a multistep process. When dsRNA is introduced into cells, it is first recognized and processed into 21–23 base-pair small interfering RNAs (siRNA) by Dicer, a RNase III family ribonuclease. These short interfering RNAs are then incorporated into and direct the RNA-induced silencing complex (RISC) to the target RNA. RISC is a nuclease complex that is responsible for the ultimate destruction of the target RNA and gene silencing [2]. In 2001, Tuschl and colleagues [3] observed that transfection of synthetic 21 base-pair siRNA duplexes into mammalian cells

effectively silences endogenous gene expression in a sequence-specific manner. This finding heralded the use of siRNA for gene silencing in mammalian systems.

siRNA oligonucleotides (21–22 base pairs) can be generated by chemical synthesis [4] or by in vitro transcription using T7 RNA polymerase [5]. Alternatively, siRNAs can be endogenously expressed in the form of short hairpin RNA (shRNA), delivered to cells via plasmids or viral/bacterial vectors [6]. Chemically synthesized siRNAs are relatively simple and quick to generate. In recent years, a number of commercial manufacturers have started to offer siRNA oligonucleotide synthesis, which has greatly facilitated the use of synthetic siRNAs in research. In this chapter, I will focus on procedures that utilize commercially synthesized siRNAs to knockdown gene expression in mammalian cells.

2 Materials

2.1 siRNA Oligonucleotides

1. Gene-specific siRNA oligonucleotides (*see Note 1*): siRNA sequences can be designed using freely available online tools and then custom-synthesized by commercial vendors (e.g., Integrated DNA Technologies or Thermo Fisher Scientific Inc.). Alternatively, pre-designed or validated siRNA oligonucleotides for specific genes can be purchased from manufacturers (e.g., GE Healthcare Dharmacon Inc., QIAGEN, or Thermo Fisher Scientific Inc.).
2. Negative control (scrambled or non-targeting) siRNA oligonucleotides
Non-targeting siRNA oligonucleotides (*see Note 2*) are siRNAs that lack complementary RNA sequences in the targeting genome. These siRNAs serve as negative controls. They can be purchased from siRNA oligonucleotide manufacturers (e.g., GE Healthcare Dharmacon Inc., QIAGEN, or Thermo Fisher Scientific).
3. Positive control siRNA oligonucleotides
Positive control siRNA oligonucleotides (*see Note 3*) are siRNAs known to downregulate the expression of a specific gene.

2.2 Cell Culture Reagents

1. Mammalian cells.
Cells to be used to perform siRNA knockdown (*see Note 4*). Here, the pancreatic cancer cell line MIA PaCa-2 is used as an example.
2. Cell culture medium appropriate for the cells being cultured.
For MIA PaCa-2, we use RPMI-1640 supplemented with 10% heat-inactivated fetal bovine serum (FBS) (*see Note 5*).

3. Phosphate Buffered Saline (PBS), pH 7.4.
4. Gibco[®] Trypsin–EDTA solution.

This is a ready-to-use trypsin solution containing 0.025% trypsin and 0.01% EDTA in PBS.

2.3 Transfection Reagents

1. siLenFect[™] Lipid Reagent (Bio-Rad Laboratories Inc) (*see Note 6*).
2. Amaxa Nucleofector[™] Kit V (Lonza Cologne AG) (*see Note 7*).

2.4 Other Reagents/ Equipment

1. RNase-free water.
2. RNA Oligonucleotide Annealing Buffer (5×) (*see Note 8*).
Potassium Acetate: 100 mM.
HEPES–KOH: 30 mM, pH 7.4.
Magnesium Acetate: 2 mM
3. Nucleofector[™] Device (Lonza Cologne AG).

3 Methods

3.1 Design of Gene-Specific siRNA Sequences

1. Designing a highly effective and specific siRNA sequence is the first step for successful knockdown of a target gene. Various groups have developed specific guidelines for designing siRNAs [7–9] and a number of online design tools are freely available (e.g., <http://sirna.wi.mit.edu>; <https://rnaidesigner.thermofisher.com/rnaiexpress>; and <http://dharmacon.gelifsciences.com/design-center>).
2. Several siRNA manufacturers such as Thermo Fisher Scientific, QIAGEN, and GE Dharmacon have predesigned (and sometimes validated) siRNA sequences for most known genes in the human genome. Researchers only need to enter the gene name or ID online to order siRNA oligonucleotides specifically designed to target the gene of interest (*see Note 9*).

3.2 Preparation of siRNA Solution

1. siRNA oligonucleotides purchased from commercial vendors (e.g., QIAGEN and Thermo Fisher Scientific) are usually ready-to-use duplex RNAs and do not need to be desalted or annealed. Simply resuspend the lyophilized siRNA duplex powder in RNase-free water to a final concentration of 20 μM. However, if the RNA oligonucleotides come as single-stranded RNA, then an annealing step is needed.
2. To anneal single-stranded RNA, resuspend the lyophilized siRNA powder received from the vendor in RNase-free water at a final concentration of 100 μM. Mix the solution well by pipetting up and down a few times. Aliquot the solution into

new tubes in small volumes (e.g., 20 μ L) and store at -20°C , if not to be used immediately. Combine 20 μ L of each complementary single-stranded siRNA oligonucleotides, 20 μ L of $5\times$ Annealing Buffer, and 40 μ L RNase-free water. Mix the solution by pipetting up and down a few times. Incubate the solution at 90°C for 2 min, and then slowly cool to room temperature by placing the tube in a large beaker containing room temperature water for about 1 h. Briefly centrifuge the tube to bring down all droplets from the sides and lid of the tube. The final concentration of the annealed siRNA duplex is 20 μ M.

3. Aliquot the resuspended/annealed siRNA into new tubes and store at -20°C . Do not freeze-thaw siRNA solution more than five times.

3.3 Delivery of siRNA into the Cells

Two methods have been widely used to deliver chemically synthesized oligonucleotides into mammalian cells: transfection and electroporation. Transfection uses a lipid carrier to facilitate the cellular uptake of siRNA. Electroporation uses powerful electric pulses to generate transient hydrophilic pores on the cell membrane and by doing so, allows the uptake of macromolecules, such as siRNA oligonucleotides. Here I describe the general procedures for performing these two methods (*see Note 10*).

3.3.1 Transfection Using siLenFect™

A number of lipid carriers (transfection reagents) specifically developed for siRNA oligonucleotides are commercially available. They often have different delivery efficiency in different cell types. Choosing the optimal transfection reagents for the cell type of interest may necessitate comparing reagents from different vendors. Transfection protocols vary from reagent to reagent and often require optimization for different cell types (and sometimes even different cell lines of the same type). In general, the manufacturer's recommended procedures should be used as a starting point for optimization. Here, I describe the procedures for siLenFect™ from BioRad Laboratory as a guide. The procedures are based on the instruction manual provided by the manufacturer with some modifications.

1. Grow MIA PaCa-2 cells in a T75 cell culture flask to 70–90% confluency. Wash the cells with 5 mL PBS twice. Add 1 mL of trypsin solution to the cells and incubate in a humidified CO_2 incubator for 5 min. Stop trypsinization by adding 10 mL of cell growth medium (RPMI-1640) containing 10% FBS. Transfer the cells and the media into a 15 mL conical tube and centrifuge the tube at $200\times g$ for 5 min to pellet the cells. Wash cell pellets twice with 5 mL PBS, and then resuspend cells in 10 mL serum-containing growth media. Count cells in a cell counter (e.g., the Cellometer by Nexcelom Bioscience).

2. Seed 0.5 to 1×10^6 cells / flask (*see Note 11*) in 5 mL growth media containing 10% FBS in T25 cell culture flasks (*see Note 12*). Allow cells to grow overnight at 37°C in a humidified 5% CO_2 incubator.
3. On the second day, 15–60 min before transfection, aspirate medium from the flask and add 2.5 mL fresh serum-containing growth medium to the cells.
4. For each T25 flask to be transfected, prepare 250 μL of transfection reagent solution in a 1.5 mL Eppendorf tube by adding 7.5 μL of siLenFect™ to 242.5 μL of serum-free medium (*see Note 13*).
5. For each T25 flask to be transfected, prepare 120 nM siRNA solution in 250 μL serum-free medium in a 1.5 mL Eppendorf tube (*see Note 14*). This can be done by first diluting the stock siRNA from 20 to 1 μM using serum-free medium (e.g., 5 μL of 20 μM siRNA plus 95 μL medium), and then further diluting it to 120 nM by taking 30 μL of the diluted siRNA (1 μM) and adding to 200 μL of cell-free medium (*see Note 15*).
6. Add the siRNA solution to the diluted siLenFect™ solution (*see Note 16*). Mix by tapping the tube or pipetting up and down. Incubate the mixed solution for 20 min at room temperature.
7. Add 500 μL of the siRNA/siLenfect™ complexes to the cells. Mix by rocking the flasks back and forth several times. Incubate the cells at 37°C in a humidified 5% CO_2 incubator.
8. Twenty-four to seventy-two hours following transfection, harvest cells by trypsinization as described above (*see Note 17*) to assess knockdown efficiency or examine functional effects of gene knockdown.

3.3.2 siRNA Delivery Using Electroporation

Lipid-based transfection methods work efficiently for many cell lines. However, for some cell lines and cell types, particularly primary cells and suspension cells, these methods yield low efficiency. For those hard-to-transfect cells, electroporation-based methods are often used to deliver nucleic acids. However, electroporation can induce high cell mortality, and often requires careful optimization of electroporation parameters (voltage, electric pulse length, and pulse number) to achieve high efficiency and low cell mortality. Amaxa's Nucleofector™ (Lonza Cologne AG) technology is an advanced electroporation technology that has been widely used for delivery of siRNA and other nucleic acids to hard-to-transfect cells. The company has developed an extensive database of cell type-specific electroporation programs and solutions, which has minimized the optimization process for end users. Here I describe the general protocol for using the Amaxa Nucleofector™ device and kit to deliver siRNA, using MIA PaCa-2 cells as an example. This

protocol is modified from the manual provided by the kit and device manufacturer (Lonza Cologne AG).

1. Growth MIA PaCa-2 cells in T75 cell culture flasks as described above.
2. On the day of transfection, preincubate 6-well plates containing 1.5 mL/well of serum-containing media at 37 °C in a humidified CO₂ incubator.
3. Harvest cells by trypsinization and count cells as described above.
4. Transfer cells to 15 mL conical tubes (1×10^6 cells per tube) and centrifuge at $100 \times g$ for 10 min at room temperature. Remove media by aspiration.
5. Resuspend the cells carefully in 100 μ L room temperature Nucleofector[®] Solution V per sample. Do not leave the cells in the Nucleofector[®] Solution longer than 15 min (*see Note 18*).
6. Add 1.5 μ L of siRNA (20 μ M) to the cell suspension (for a final siRNA concentration of 300 nM) (*see Note 19*). Mix by pipetting up and down.
7. Transfer cell/siRNA mixture into a certified cuvette (included in the Nucleofector[™] kit). Make sure the solution covers the bottom of the cuvette. Close the cuvette with the cap.
8. Insert the cuvette containing the cell/siRNA solution into the Nucleofector[®] Cuvette Holder. Select the Nucleofector[®] Program T-020 (*see Note 20*) and apply the program.
9. Remove the cuvette from the holder once the program is finished. Add 500 μ L of the preequilibrated culture media to the cuvette. Gently mix and transfer the solution to the preincubated 6-well plate. Use the pipettes supplied by the kit and avoid repeated aspiration of the solutions.
10. Incubate the cells at 37 °C in a humidified 5% CO₂ incubator.
11. Assess knockdown efficiency or examine functional effects of the knockdown 24–72 h following electroporation.

3.4 Assessment of Gene Knockdown Using Reverse Transcription Polymerase Chain Reaction (RT-PCR) and Western Blotting

Because siRNA oligonucleotides target mRNA for degradation, RT-PCR can be used to measure effects on gene expression using negative control siRNA-treated cells and gene-specific, siRNA-treated cells. Readers are referred to other literature for RT-PCR protocols [10–12].

Although reduction in transcript expression usually results in decreased protein abundance, mRNA levels do not always correlate with protein levels. For example, mRNA measurement can overestimate knockdown of genes whose protein products have long half-lives. Therefore, it is necessary to assess protein levels to ensure

efficient knockdown of gene expression and to determine the optimal time point for assessing cellular effects of siRNA knockdown. Western blotting is the most widely used technique for detecting proteins (*see Note 21*). Protocols for Western blotting can be readily found in the literature; for example, “Western Blotting: A guide to current methods” (edited by Hicklin T, 2015), found in a supplement to *Science* magazine.

3.5 Examination of the Functional Effects of siRNA Knockdown

Once knockdown of the siRNA-targeted gene is confirmed, assays can then be carried out to investigate resulting functional effects. Depending on the known or predicted functions of the target gene, a variety of assays (cell growth and survival, migration, apoptosis, effects on downstream signaling, etc.) can be used.

4 Notes

1. siRNA oligonucleotides designed to target different regions of a gene can have different knockdown efficiencies [13]. Although the current siRNA design algorithms are getting better at selecting efficient siRNA sequences, only about one in four siRNAs produces a knockdown efficiency of >80%. Therefore, it is imperative that multiple (usually 2–4) siRNA sequences for each target gene are obtained and optimized individually. Alternatively, multiple siRNAs can be pooled and used in a single transfection (e.g., GE Dharmacon offers pre-designed/pooled siRNA for human and other species).
2. A negative control siRNA is included in the experiment to distinguish sequence (or gene)-specific effects from non-sequence specific effects in the siRNA-treated cells. The negative control siRNAs can be siRNA sequences that have the same nucleotide composition as the gene-specific siRNA, but lack significant sequence homology to the human genome or siRNAs that have been designed to have no known homology to the human genome (often called non-targeting siRNA). Non-targeting siRNAs are commercially available from a variety of vendors (e.g., QIAGEN or GE Dharmacon).
3. Positive control siRNAs are used to monitor the efficiency of siRNA delivery to cells. These are siRNA sequences known to induce reproducible knockdown of a gene *in vitro*. If the gene targeted by the positive control siRNA is essential for cell survival then knockdown of that particular gene will result in rapid cell death, and the efficiency of siRNA delivery can be evaluated under a microscope. We have found that siRNAs targeting the Ubiquitin B (UBB) gene or the AllStars Cell Death Control siRNA from QIAGEN are good positive controls that produce rapid cell death.

4. Cells should be evaluated for the expression level of the gene of interest. To optimize siRNA knockdown conditions, a cell line expressing relatively high levels of the target gene should be used.
5. The antibiotics penicillin and streptomycin are often added to culture medium to prevent bacterial contamination. However, because transfection reagents increase cell permeability, the delivery of antibiotics may also be increased, which could result in increased cytotoxicity. Therefore, adding antibiotics to the transfection medium is not recommended.
6. A number of lipid-based transfection reagents are commercially available. Their delivery efficiency varies and can be cell line-dependent. It is advisable to first consult the literature to determine if any other groups have reported siRNA transfection in the same cell lines/types, and then start with the same transfection reagents and conditions.
7. Lonza has developed five different Nucleofector™ Solutions designed to work for different cell lines/types. The manufacturer has also developed optimized protocols for a number of cell lines and primary cell types. Kits containing the optimized Nucleofector™ Solution recommended by the manufacturer should be purchased. If the cell line or type is not on the list with an optimized protocol, then an optimization kit should be obtained.
8. Other annealing buffers have also been reported in the literature, e.g., 50 mM Tris, pH 7.5–8.0, 100 mM NaCl, and 5 mM EDTA (5×).
9. Two to four siRNA sequences that target different regions of a gene of interest should be tested (*see Note 1*).
10. The two delivery methods have their own advantages and disadvantages [14]. The transfection method is simple and requires no specialized equipment, but it does not work well with primary cells and suspension cells. The electroporation method can achieve very high delivery efficiency, even in hard-to-transfect cells, although it often causes high cell death. It also requires specialized equipment (i.e., an electroporator). Selecting the right method will depend on the experimental conditions, such as the cells being used and the assays to be run after transfection.
11. The exact cell number to seed must be optimized for different cell lines. The goal is to achieve 50–70% confluency on the following day.
12. Depending on the assays to be run after transfection, other culture vessels can be used. Depending on the surface area of the vessel, the amount of reagents may need to be scaled up or

down. Refer to the manufacturer's manual for recommended medium volumes and the amount of reagents for different culture vessels.

13. The amount of siLenfect™ may need to be optimized using a range of volumes from 2.5 to 20 μ L.
14. The concentration of siRNA needed for efficient knockdown may vary depending on cell lines used and the gene target itself. It is advisable to optimize the concentration of siRNA by carrying out transfections using different siRNA concentrations ranging from 5 to 20 nM (final concentration).
15. Master mix can be prepared if replicates of the same siRNA concentrations are being carried out.
16. It is recommended that a mock transfection with only siLenfect™ (no siRNA added) be included as a control.
17. Gene knockdown can be detected as early as 4 h and could last up to 5 days, and even 7 days in some cases [15]. However, in general, 24–96 h is the ideal time periods for accessing gene knockdown and investigating functional effects of the siRNA knockdown in cell culture. Cells can be retransfected with the siRNA to extend the duration of gene knockdown.
18. Do not leave cells in the Nucleofector® Solution for longer than 15 min, as longer exposure may lead to reduced transfection efficiency and cell viability.
19. The optimal siRNA concentration may vary from cell line to cell line. A range (30–300 nM) of concentrations should be used if an optimal concentration is not known.
20. The manufacturer has optimized programs for a number of cell lines. Please refer to the manufacturer's website for details (<http://www.lonza.com/research/>).
21. In addition to Western blotting, other methods such as immunofluorescence staining and ELISA (enzyme-linked immunosorbent assay) can be used to monitor the knockdown of gene expression by siRNA.

Acknowledgments

This work was supported by NIH/NCI grants CA169281 and CA191923 to H.H.

References

1. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE et al (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391:806–811
2. Sontheimer EJ (2005) Assembly and function of RNA silencing complexes. *Nat Rev Mol Cell Biol* 6:127–138
3. Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K et al (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* 411:494–498
4. Micura R (2002) Small interfering RNAs and their chemical synthesis. *Angew Chem Int Ed Engl* 41:2265–2269
5. JY Y, DeRuiter SL, Turner DL (2002) RNA interference by expression of short-interfering RNAs and hairpin RNAs in mammalian cells. *Proc Natl Acad Sci U S A* 99:6047–6052
6. Brummelkamp TR, Bernards R, Agami R (2002) A system for stable expression of short interfering RNAs in mammalian cells. *Science* 296:550–553
7. Birmingham A, Anderson E, Sullivan K, Reynolds A, Boese Q et al (2007) A protocol for designing siRNAs with high functionality and specificity. *Nat Protoc* 2:2068–2078
8. Molecule of the month. Y-27632 *Drug News Perspect* 14:45
9. Park YK, Park SM, Choi YC, Lee D, Won M et al (2008) AsiDesigner: exon-based siRNA design server considering alternative splicing. *Nucleic Acids Res* 36:W97–W103
10. Guan H, Yang K (2008) RNA isolation and real-time quantitative RT-PCR. *Methods Mol Biol* 456:259–270
11. Tsai SJ, Wiltbank MC (1996) Quantification of mRNA using competitive RT-PCR with standard-curve methodology. *BioTechniques* 21:862–866
12. Schmittgen TD, Livak KJ (2008) Analyzing real-time PCR data by the comparative C (T) method. *Nat Protoc* 3:1101–1108
13. Dykxhoorn DM, Novina CD, Sharp PA (2003) Killing the messenger: short RNAs that silence gene expression. *Nat Rev Mol Cell Biol* 4:457–467
14. Gilmore IR, Fox SP, Hollins AJ, Akhtar S (2006) Delivery strategies for siRNA-mediated gene silencing. *Curr Drug Deliv* 3:147–155
15. Dorsett Y, Tuschl T (2004) siRNAs: applications in functional genomics and potential as therapeutics. *Nat Rev Drug Discov* 3:318–329

Using Luciferase Reporter Assays to Identify Functional Variants at Disease-Associated Loci

Anup K. Nair and Leslie J. Baier

Abstract

The genomic era, highlighted by large scale, genome-wide association studies (GWAS) for both common and rare diseases, have identified hundreds of disease-associated variants. However, most of these variants are not disease causing, but instead only provide information about a potential proximal functional variant through linkage disequilibrium. It is critical that these functional variants be identified, so that their role in disease risk can be ascertained. Luciferase assays are an invaluable tool for identifying and characterizing functional variants, allowing investigations of gene expression, intracellular signaling, transcription factors, receptor activity, and protein folding. In this chapter, we provide an overview of the different ways that luciferase assays can be used to validate functionality of a variant.

Key words Dual-luciferase assay, Functional variant, GWAS, Firefly luciferase, Renilla luciferase

1 Introduction

The GWAS study design has been extremely successful in identifying genetic variants associated with both common and rare diseases [1, 2]. The main advantage of GWAS is that not all variants in the genome need to be directly genotyped due to linkage disequilibrium (LD) among nearby variants. Because of LD with a causal variant, a marker variant would show association with the disease of interest [1, 2]. While LD provides a way to genotype a small number of markers to capture the effect of many variants, it obscures which variant is the actual functional basis of the observed association. While many studies are moving away from GWAS due to decreasing costs of massively parallel, whole genome or exome sequencing [3–5], LD still remains a problem with these approaches. Also, it is possible that a single variant may regulate multiple genes at the same locus through a *cis*-acting effect or at a different locus via a *trans*-acting effect [6–8]. Historically, it was assumed that functional variants would lie within coding regions, and therefore, affect the structure or function of the protein product. However, the vast

majority of disease-associated variants map to intronic or intergenic regions [9]. Given the observations from large studies like Encyclopedia of DNA Elements (ENCODE) and the National Institutes of Health Roadmap Epigenomics projects, which suggest that a major part of the human genome is functional [10, 11], it can be inferred that many common diseases result from abnormal gene regulation, as opposed to structural or functional abnormalities in the protein product. Identifying the causal mutation from numerous proximal variants all showing association with the disease phenotype due to LD must be accomplished by functional characterization in an *in vitro* system. To date, very few large scale and massively parallel *in vitro* assays have been undertaken to functionally characterize the large amount of data obtained from GWAS, WGS, and WES, to achieve clinically meaningful observations. Here, we describe the various ways by which luciferase assays can be used to identify disease-associated loci.

1.1 Luciferase Assay

A luciferase assay is a type of reporter gene assay used to study intracellular signaling, gene expression, receptor activity, transcription factors, mRNA processing, and protein folding. In its simplest form, a regulatory element, for example, a promoter region, is cloned upstream of a reporter gene, such as *luciferase*, in an expression vector, which is a plasmid that will express the reporter. This construct (i.e., expression vector + reporter gene + regulatory element) is transfected into an appropriate cell line. Once inside the cell, the regulatory element utilizes the transcriptional machinery of the cell to express the reporter. The cells are then assayed for the presence of the reporter itself, or the enzymatic activity of the reporter, which directly correlates with the activity of the regulatory element (Fig. 1). The regulatory element used in the assay can be manipulated to contain either the reference allele or an alternate allele of the variant of interest. A difference in activity between the two forms of the regulatory element is an indicator of the functional impact of the variant. A good reporter is easily identified and measured quantitatively when expressed inside the cells. One of the most commonly used reporters is the luciferase gene, which produces the luciferase enzyme and can be quantitatively measured by a bioluminescence assay with high sensitivity [12].

Bioluminescence-based assays are ideal in many respects. They can be measured instantaneously, are highly sensitive, and have a wide dynamic range; yet do not have endogenous activity, which would interfere with quantification in cells. The bioluminescence assays used to quantify luciferase enzyme activity utilize its interaction with a bioluminescent substrate (luciferin) to produce light. The emitted light can be measured with a luminometer. Although different luciferase reporter genes are available, the two most commonly used ones are from firefly (*Photinus pyralis*) and *Renilla* (*Renilla reniformis*). The luciferases from firefly and *Renilla* have

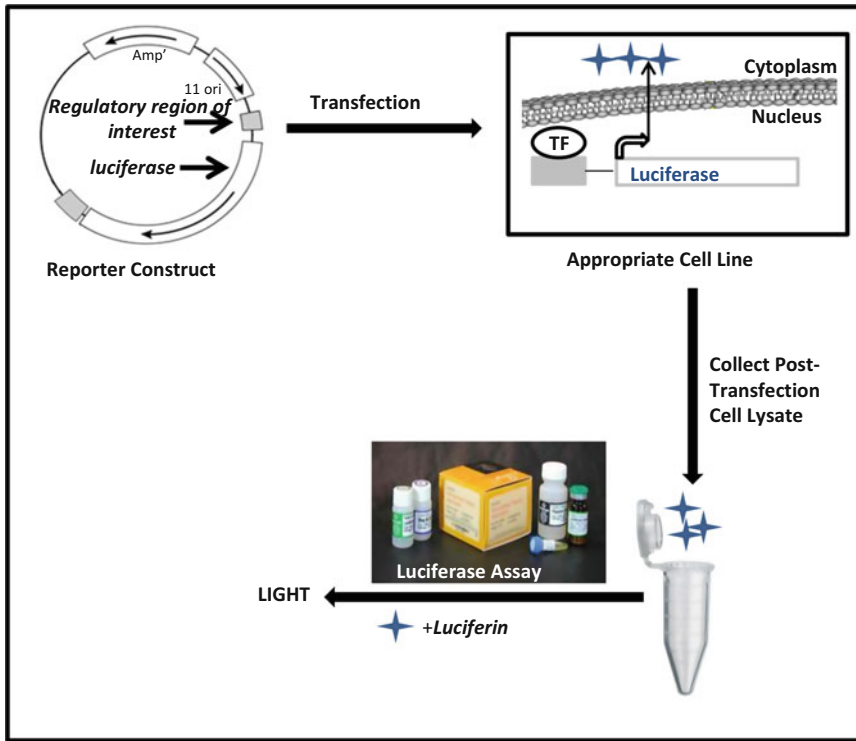


Fig. 1 General overview of the luciferase assay. Transient transfection of a Luciferase promoter–reporter construct results in production of Luciferase enzyme. The post-transfection cell lysate containing the Luciferase enzyme is collected and assayed. The light produced by the luciferase assay can be measured using a luminometer, and is proportional to the activity of the promoter. *TF* = transcription factor

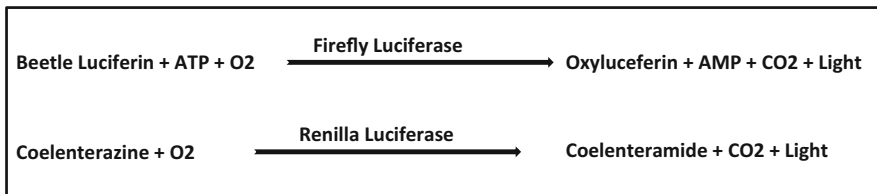


Fig. 2 Chemistry involved in luciferase assay. Beetle Luciferin acts as the substrate for firefly luciferase and coelenterazine is the substrate for *Renilla* luciferase. Both reactions result in the production of light that can be measured using a luminometer

different enzyme structures and utilize different substrates, due to distinct evolutionary origins [13]. Neither firefly (61 kDa), nor *Renilla* (36 kDa) luciferase require post-translational processing for enzyme activity, and can therefore act as genetic reporters immediately upon translation [14, 15]. The chemistry involved in the assay is shown in Fig. 2.

1.2 Dual-Luciferase Assay

Due to the inherent nature of biological experiments, many unintended variables (e.g., pipetting inaccuracy, differences in transfection efficiency, variation in cell density, etc.) may be introduced

during a luciferase assay, which may confound results. Like real-time PCR, where normalization with an endogenous control is necessary to obtain trustworthy results, here also normalization with an internal control will provide greater confidence in the observed results. The different chemistry and substrate utilized by firefly luciferase and *Renilla* luciferase provide an ideal system to use the two luciferase reporter genes in a single assay, which are referred to as dual-luciferase assays (DLR assay). While the luminescent signals provided by both luciferases have similar sensitivities, the firefly luciferase chemistry produces a flash of light that decays rapidly after mixing of substrate and enzyme. In contrast, the *Renilla* luciferase reaction provides a signal that decays slowly over the course of the measurement [16]. In a DLR assay, either the firefly luciferase or *Renilla* luciferase can be used as the experimental reporter gene, while the other serves as the control reporter. The luminescent signal produced by the control reporter, which should be similar, theoretically, in all experiments, is then used to normalize the signal from the experimental reporter. In most cases, the firefly luciferase is used as the experimental reporter and *Renilla* luciferase is used as the control reporter. The dual-luciferase assay kit from Promega, for example, provides a fast and convenient option for performing DLR assays (*see Note 1*).

A drawback to using the *Renilla* luciferase is the low-level autoluminescence emitted by Coelentrastazine, the substrate of *Renilla* luciferase, which is exacerbated in the presence of nonionic detergents (e.g., Triton X-100) used for cell lysis. However, the DLR assay system from Promega offers a proprietary chemistry that reduces autoluminescence to undetectable levels with most luminometers. Likewise, there is a potential for *trans*-effects between promoters on cotransfected plasmids in the DLR assay. These *trans*-effects can affect reporter gene expression independent of the experimental conditions, thereby, confounding results. In most cases, these effects are seen when the experimental, the control, or both reporters contain strong promoter/enhancer elements. A *trans*-effect can be avoided by using very low amounts of the control reporter vector, just enough to maintain a low constitutive expression of the control luciferase. Additionally, many different factors have been identified that influence the expression of *Renilla* luciferase, and these need to be taken into consideration when designing an experiment [17].

1.3 Luciferase Assay to Identify Disease-Associated Locus

As noted, large association studies have identified hundreds of variants associated with common diseases. However, in most cases, the true functional variant remains unknown. Also, because the majority of associated variants are either intronic or intergenic, and not in LD with a coding variant, it is likely that potential functional variants lead to disease via regulatory effects on gene expression. Luciferase assays are useful for identifying potential

functional variants. In the following section, we briefly outline how luciferase assays can be used to identify functional variants that affect gene expression and to study the functionality of coding variations.

1.4 Cloning Strategy

Depending on the genomic location of the variants, different cloning strategies can be utilized for luciferase assays. The following section describes some of the strategies.

1.4.1 Assessment of Noncoding Variants

1. *Selecting a gene whose promoter will be assayed (see Note 2)*

While the majority of variants associated with common disease are noncoding, and in some cases, it is not even apparent which gene is contributing to disease susceptibility, knowledge of the potential target gene is helpful in the design of luciferase assays. Having a gene in hand allows the effect of the variant to be tested on its minimal promoter and provides a strong rationale for the selection of the appropriate cell type (e.g., one containing the required “transcriptional machinery”) for the luciferase assay.

There are several databases that can aid in physiologically or mechanistically connecting a variant to a specific gene. For example, some variants are known to function as *cis*-acting expression quantitative trait loci (eQTLs) for specific genes. The Genotype-Tissue Expression (GTEx) project consortium has produced publicly available expression data in different tissue types that can be used to assess whether the disease-associated variant is a known *cis*-eQTL for a gene [18]. Additionally, epigenetic data can be used to verify whether a disease-associated variant is located within a region with promoter or enhancer activity, or if the variant has possible regulatory effects using available computational web tools like HaploReg and RegulomeDb [19, 20]. Finally, literature searches can be performed for all nearby genes to assess whether one of these genes has been implicated in the pathogenesis of the disease of interest. A list of computational tools for identifying target genes or prioritizing variants for functional studies can be found elsewhere [7].

2. *Cloning of the selected promoter and regulatory element*

Once a potential target gene is identified, the minimal promoter of that gene can be cloned in the multiple cloning site [MCS], upstream of the luciferase gene in a promoter-less luciferase vector (i.e., pGL3 basic vector or pGL4.10[*luc2*]). The restriction enzymes used for cloning should allow for additional restriction sites upstream of the inserted minimal promoter to facilitate further application. This minimal promoter–reporter construct can be used to standardize transfection and luciferase assay conditions in a cell type in which the promoter is known to be active. If the promoter is active in the chosen cell type, it will lead to the production of luciferase enzyme, which can be assayed using a luciferase assay (Fig. 3).

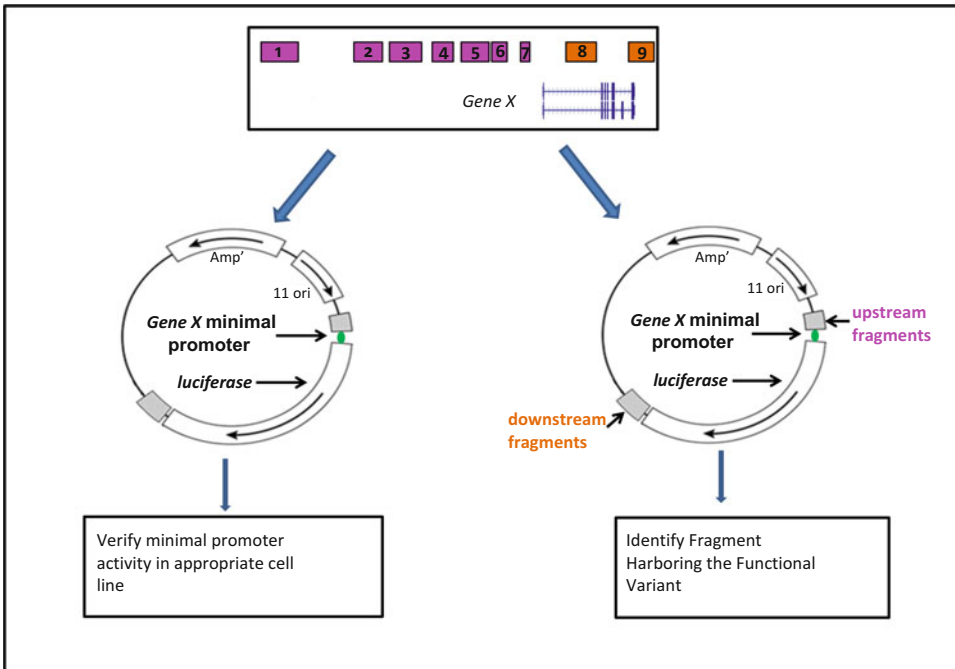


Fig. 3 Cloning strategy to identify noncoding functional variants. The disease-associated variant in this example is upstream of a gene with a priori physiologic relevance to this disease (Gene X, see the “selecting a gene whose promoter will be assayed” section). The variant is in LD with variants both upstream and downstream of the target gene. The region is divided into fragments and cloned upstream or downstream of the minimal promoter–reporter construct. Each fragment contains SNPs in LD with the variant. The fragments are assayed for differences in luciferase activity between constructs containing the reference and alternate alleles. The fragment with differences in activity is further studied to identify the functional variant

The cloning of the regulatory element can be done in multiple ways, depending on the number and proximity of all associated variants, due to LD structure in that region. The most simplistic, albeit rare, scenario is an associated variant positioned in a known promoter region as a singleton (i.e., not in LD with other variants). In this case, the promoter fragment harboring the variant can be cloned upstream of the luciferase gene in the pGL3 basic vector or pGL4.10[*Luc2*] vector. Two constructs must be made, one with the reference allele of the variant, and the other containing the alternate allele. These constructs can then be assayed individually to analyze the functionality of the variant. If there is more than one variant of interest in the promoter, site-directed mutagenesis can be used to make the following constructs, with the reference allele at all the variants of interest and alternate alleles at individual variants. It is preferable to assay all constructs in the same experimental setup.

The more common situation, however, is the presence of multiple variants in a region not known to have promoter activity, which all associate with the disease phenotype. In this scenario, if the associated variants are located close to each other, then a

fragment containing all the variants can be assayed by cloning it either upstream of the minimal promoter using the available restriction sites in MCS of the minimal promoter–reporter construct, or downstream of the luciferase gene using the restriction sites for the BamHI and SalI enzymes. Site-directed mutagenesis can then be used to create constructs containing individual SNPs for additional assays to identify the functional variant. If the number of associated variants spans a large region, then variants can be prioritized for assays based on computational assessment of regions with potential regulatory activity in disease-relevant cell type. If all associated variants in the region are to be assayed, then the region can be divided into multiple fragments of approximately equal size. Each fragment should contain either the reference allele or the alternate allele of all associated variants. These fragments can then be cloned either upstream of the minimal promoter–reporter construct, or downstream of the luciferase gene, depending on the actual genomic organization (Fig. 3). Doing so will minimize the number of potential functional variants to be individually assayed. Then, the construct showing the largest difference in activity can be used as a template for site-directed mutagenesis to create additional constructs differing at only one variant. These constructs can subsequently be assayed to assess the functionality of each individual variant.

1.4.2 Cloning Strategy: Assessment of Coding Variants

Different cloning strategies can be used to functionally assess a coding variant using luciferase assay, depending on the function of the protein. Two examples of the utility of luciferase assays to assay coding variants that either directly or indirectly affect transcription factors are provided below. However, for coding variants within genes whose protein products are not involved in regulating transcription, a different assay should be used to assess functionality.

For a coding variant in a gene encoding a transcription factor (TF), the promoter of a known target gene that has the binding sites for the TF, or a synthetic construct with multiple sites for the same TF upstream of the minimal promoter of a responsive gene, can be cloned into the MCS of the pGL3 basic or pGL4.10[*Luc2*] vector. This vector serves as the promoter–reporter construct. The gene (cDNA) coding for the TF can then be cloned in-frame into a mammalian expression vector (e.g., pCDNA3.1 or pCMV6-AC). Alternatively, a full-length cDNA clone or a full-length open reading frame clone in a mammalian expression vector can be obtained from commercial sources (e.g., Origene). Site-directed mutagenesis can be used to create the desired mutation. This serves as the experimental construct: one with the reference allele of the coding variant and one with the alternate allele. The promoter–reporter construct and the experimental construct is then

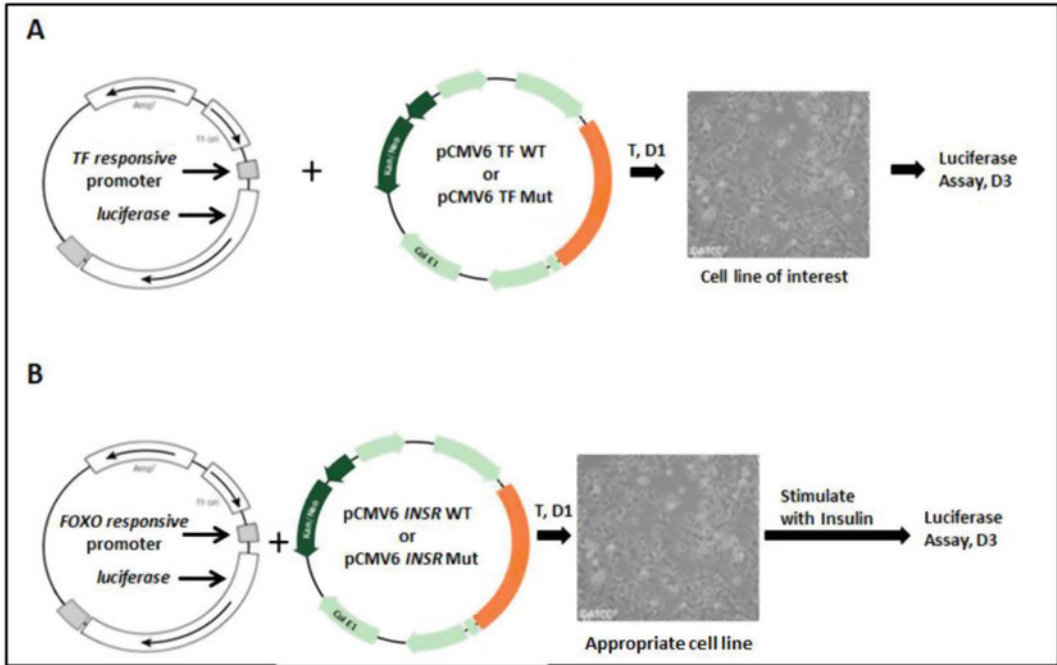


Fig. 4 Cloning strategy to functionally study coding variation using a luciferase assay. **(a)** Mutations in transcription factors. An expression vector containing the full-length ORF of the TF (either with the reference or alternate allele of the mutation of interest) is cotransfected with a luciferase vector containing the response element for the TF upstream of the luciferase gene. **(b)** Mutation in the insulin receptor gene. An expression vector containing the full-length ORF of the insulin receptor (either with the reference or alternate allele of the mutation of interest) is cotransfected with a luciferase vector containing the FOXO response element (Insulin response element) upstream of the luciferase gene in an appropriate cell line. One day post-transfection, the cells are stimulated with insulin to activate the insulin signal transduction pathway, which is followed by a luciferase assay to measure activity. If the cell line used expresses the insulin receptor endogenously, consistent change in luciferase activity should be observed after stimulation with insulin, when overexpressing the insulin receptor along with the response element compared to overexpression of the response element alone. The assay can also be done using different concentrations of insulin, and then analyzing the ability of the insulin receptor (mutant vs. WT) to inhibit FOXO-mediated transcription in a dose-dependent (insulin) manner. *TF* = transcription factor, *T*—transfection, D1–3—days 1, 2, and 3

cotransfected into a suitable cell line, along with a *Renilla* luciferase vector to serve as the internal control for normalizing transfection efficiency, and assayed by the dual-luciferase assay (Fig. 4a).

For a coding variant that affects a signal transduction pathway, a promoter containing multiple sites responsive to the end effector TF of that pathway can be used for the assay. For example, to study a mutation in the insulin receptor protein, the FOXO/insulin signal transduction pathway can be used. Signaling through the insulin receptor is known to inhibit transcription mediated by FOXO TFs. A mutation in the insulin receptor can be studied using a FOXO responsive promoter (insulin responsive element) as the insert in the promoter-reporter construct. The

promoter–reporter construct can then be cotransfected with the insulin receptor expression construct containing either the reference or variant allele (experimental construct). Following transfection, the cells can be stimulated by insulin and inhibition of FOXO activity can be measured as a function of decrease in luciferase activity (Fig. 4b).

**1.4.3 Cloning Strategy:
Assessment of Variants in a
3'-UTR**

A 3'-UTR is the region of mRNA immediately following the translational termination codon. This region often contains elements that post-transcriptionally regulate gene expression through mechanisms including polyadenylation, translation efficiency, localization, or stability. The 3'-UTR can also harbor binding sites for micro-RNAs (miRNAs) and regulatory proteins. miRNAs bind the 3'-UTR and downregulate gene expression by translational inhibition or transcript degradation [21]. Likewise, regulatory proteins bind to the 3'-UTR and in most cases, repress transcript expression. Variants that create or destroy miRNA or regulatory protein binding sites can therefore modulate gene expression and potentially affect disease conditions. Disease-associated variants affecting miRNA binding have been described [22].

To analyze the function of these variants, the region of interest (containing either reference or alternate allele) is cloned downstream of the luciferase gene. The pmirGLO vector, which has a MCS immediately after the translation termination codon of the *luc2* gene, is recommended for this assay. Upon transfection, the cloned insert will serve as the 3' UTR of the luciferase gene. Theoretically, when transfected into an appropriate cell line, the 3'-UTR will regulate the luciferase gene in a manner similar to the target gene. A variant in the 3' UTR that affects its function can be then assayed by a luciferase assay by comparing the effects of the reference and alternate alleles on luciferase expression. Other vectors like psiCHECK™-1 and psiCHECK™-2, which have different features, can also be used to assay 3'-UTR variants.

2 Materials

Perform all tissue culture and transfection procedures aseptically under a cell culture hood. Store all reagents according to the manufacturer's instructions. Minimize experimental variation by preparing working stocks of all reagents and avoiding repeated freeze-thaws of stocks. If possible, use tissue culture serum from the same lot, as serum is a complex, undefined mixture whose variability among lots can affect experimental reproducibility. For optimal transfection efficiency, select cells that have undergone few passages.

2.1 Cloning

2.1.1 *Vectors*

1. pGL3 basic (Promega).
2. pGL4.10[*luc2*] (Promega).
3. pRL-TK (Promega).
4. pGL4.74[hRLuc/TK] (Promega).
5. pmirGLO vector (Promega).
6. psiCHECK™-1 and psiCHECK™-2 (Promega).
7. pCMV6-AC (Origene).
8. pCMV6-EV (Origene).

2.1.2 *cDNA Clones*

Full-length mouse or human cDNA clones, or full-length ORF clones in a mammalian expression vector, can be obtained from Origene or similar vendors. A commonly used construct is a full-length ORF clone in pCMV5 or pCMV6 vectors.

2.1.3 *Enzymes*

1. DNA Polymerase—High fidelity polymerase should be used to amplify the region of interest.
2. Restriction enzymes—Select the appropriate restriction enzyme to cleave within the multiple cloning sites (MCS) of the vectors.
3. Alkaline Phosphatase—Calf Intestinal Alkaline Phosphatase (CIP) or Shrimp Alkaline Phosphatase.
4. T4 DNA Ligase.

2.1.4 *Kits*

1. PCR purification kit.
2. Gel extraction kit.
3. Plasmid isolation kit.
4. Site-directed mutagenesis Kit.

2.1.5 *Competent Cells*

1. NEB® 5-alpha Competent E. coli (High Efficiency).

2.2 Cell Culture

1. Cell line—A cell line should be selected that expresses the transcription factors affecting the regulatory region being tested (*see Note 3*). Most cell lines can be obtained from *American Type Culture Collection (ATCC)*.
2. Trypsin-EDTA.
3. 5% CO₂ incubator.
4. 1 × sterile Phosphate Buffered Saline (PBS).
5. Tissue culture grade culture wares.
6. 15 and 50 mL centrifuge tubes and centrifuge.
7. Cell counter of choice.

2.3 Transfection

1. Lipofectamine (Lipofectamine[®] LTX with Plus[™] or Lipofectamine 3000 reagent) is commonly used for transfection. Lipofectamine 3000 provides good efficiency for transfecting most cell lines.
2. Opti-MEM[®] I Reduced Serum Medium.

2.4 Cell Harvesting

1. 1 × PBS.
2. 1 × Passive Lysis Buffer (PLB). Dilute 5 × PLB from Dual-Luciferase Assay Kit by adding 4 volumes of water to 1 volume of 5 × PLB.

2.5 Luciferase Assay (Dual-Luciferase[®] Reporter Assay System)

Luciferase Assay Substrate—Resuspend the lyophilized luciferase assay substrate using the luciferase assay buffer II provided in the Dual-Luciferase[®] Reporter Assay kit. Divide into aliquots and store aliquots at -70°C . Avoid repeated freeze-thaws. Thaw aliquots in room temperature water bath.

1 × Stop and Glo Reagent—Add 1 volume of 50 × Stop and Glo substrate to 50 volumes of Stop and Glo buffer from the Dual-Luciferase[®] Reporter Assay kit. This serves as a quencher for firefly luciferase and a substrate for the *Renilla* luciferase. 1 × Stop and Glo reagent should be freshly prepared for each assay.

3 Methods**3.1 Cloning**

1. To create experimental constructs, amplify the region harboring the associated variant using human genomic DNA as a template and High Fidelity DNA Polymerase according to the manufacturer's protocol (*see Note 4*). If available, two templates should be chosen, one from an individual homozygous for the reference allele and one from an individual homozygous for the variant allele. If a template for the variant allele is not available, site-directed mutagenesis can be used to create one. The oligonucleotides used for amplifying should be designed in such a way that they have the appropriate restriction enzyme sites at their 5' end (*see Note 5*).
2. Use a small amount (i.e., 5 μL) of the amplicon for agarose gel electrophoresis to verify size.
3. Follow the instructions in the PCR purification kit to purify the remaining amplicon.
4. Use 1 μg of the selected vector and purified amplicon to set up two separate double-digestion reactions. Set up the double-digestion in a total volume of 30 μL . Use the double-digest finder web tool from NEB to select the reaction conditions according to restriction enzyme combination.

5. Incubate the double-digestion reaction for 6–8 h. After 6 h, add 1 μL of Calf Intestinal Alkaline Phosphatase to the vector double-digest only, mix well, and incubate at 37 °C for 1 h (*see Note 6*).
6. Heat-inactivate the double-digest mix depending on the selection of the restriction enzyme.
7. Pipette the entire volume of the double-digest reaction onto an agarose gel for separation by electrophoresis.
8. Extract the double-digested amplicon and vector from the agarose gel using a gel-extraction kit according to the manufacturer's protocol.
9. Quantify the eluted products by agarose gel electrophoresis.
10. Set up the ligation reaction with T4 DNA ligase using a molar vector: insert ratio of 1:3 for sticky-end ligation and 1:6 for blunt-end ligation. Incubate for 16 h at 16 °C, followed by heat inactivation for 20 min at 65 °C (*see Note 7*).
11. Use the ligation mixture for transformation of high efficiency competent cells following the manufacturer's protocol. Incubate transformation plates for 12–14 h.
12. Screen colonies to select positive clones by PCR.
13. Inoculate positive colonies for each construct in 3 mL LB broth and incubate at 37 °C in a shaker incubator at 220 RPM for 8–10 h.
14. Isolate clones using a plasmid isolation mini kit following the manufacturer's protocol. Quantify isolated clones.

3.2 Cell Culture and Transfection (See Note 3)

1. The type of cell culture ware should be determined based on the number of constructs. It is recommended that each construct be assayed in duplicate.
2. Cells should be cultured based on the conditions recommended by the vendor. Antibiotics can be used for cell culture, but not during transfection.
3. Cells for transfections should be seeded in such a way that they will be 60–90% confluent on the day of transfection (*see Note 8*).
4. On the day of transfection, mix the experimental construct and the control vector (expressing *Renilla* luciferase) in a ratio of 10:1. For a 12-well plate, use 900 ng of the experimental construct and 100 ng of the *Renilla* luciferase vector. The ratio differs based on assay conditions and should be determined empirically. Use equal amounts of each experimental construct to be analyzed.

5. Depending on the cell line, either Lipofectamine 3000 or Lipofectamine LTX and Plus can be used for transfection. Transfections are performed per the manufacturer's protocol (*see* **Notes 9–14**).

3.3 Cell Harvesting

1. Cells can be harvested 48 h after transfection.
2. For harvesting, remove the spent media from the wells and wash cells with $1 \times$ PBS. Make sure to completely remove PBS.
3. Add enough $1 \times$ PLB to cover the cell surface; usually $250 \mu\text{L}$ /well for a 12-well plate is sufficient (*see* **Note 15**).
4. Incubate at room temperature in an orbital shaker for 15 min.
5. Collect the lysate including any debris in a 1.5 mL centrifuge tube. The lysate can be stored at -70°C .

3.4 Dual-Luciferase Assay (See Notes 16–18)

1. Bring all reagents to room temperature prior to starting the assay.
2. Thaw the cell lysate, which is stored at -70°C , to room temperature. Thoroughly vortex samples before the assay.
3. Aliquot $100 \mu\text{L}$ of LARII reagent from the dual-luciferase assay kit into luminometer glass tubes. Aliquot extra tubes as sometimes one sample may need to be measured more than once.
4. Prepare $1 \times$ Stop and Glo reagent, enough to use $100 \mu\text{L}$ per sample. Prepare extra in case of additional measurements.
5. Transfer $20 \mu\text{L}$ of cell lysate to $100 \mu\text{L}$ of LARII tube and mix well by pipetting. Immediately measure luminescence using a luminometer. Record the reading.
6. Add $100 \mu\text{L}$ of $1 \times$ Stop and Glo reagent and mix by vortexing. Immediately measure luminescence using a luminometer. Record the reading.
7. Repeat **steps 5 and 6** once more for each sample.

3.5 Data Analysis (See Notes 19 and 20)

1. Average luminescence readings from firefly luciferase and *Renilla* luciferase for each sample.
2. Calculate the relative firefly luciferase/*Renilla* luciferase activity for each sample using the average readings. This is the relative luciferase activity.
3. To calculate the fold-enrichment in relative luciferase activity of the experimental construct over the vector alone control, divide the relative luciferase reading of the experimental construct by the relative luciferase reading of the vector alone control.
4. Each experiment should include at least two experimental replicates for each construct and the experiment should be repeated at least 6–8 times.

5. Mean fold-enrichment in luciferase activity of each construct is calculated for each experiment by averaging the fold-enrichment obtained in experimental replicates.
6. An unpaired t-test can be used to calculate whether the relative luciferase activity or fold-enrichment in luciferase activity is significantly different between two experimental constructs.

In conclusion, this chapter provides an outline of several ways in which luciferase assays can be used to identify functional variants in the genome. Although general approaches to study promoter, intronic, intergenic, 3' UTR, and coding variants are described, it should be noted that the genome is highly enriched for additional regulatory elements such as long noncoding RNAs, which are not discussed here. As further knowledge of the genome becomes available, it is likely that more variation may be assessed for functionality using luciferase assays.

4 Notes

1. While this chapter describes a protocol using the dual-luciferase reporter assay system provided by Promega, other luciferase reporter systems are commercially available (e.g., Pierce Renilla-Firefly dual assay from Thermo-Fisher or luciferase assay systems from Switchgear Genomics). Many vendors also provide precloned promoter, 5'UTR, and 3'UTR reporter constructs that can be used depending on the region of interest.
2. A pGL3-promoter vector can be used for variants that are in genomic regions with no clear candidate genes and where a minimal promoter cannot be identified for the promoter-reporter construct. These vectors have a known promoter element (e.g., SV40 promoter) upstream of the MCS. Promega offers luciferase reporter vectors with different promoters upstream of the MCS, which can be used depending on the assay and cell type of interest.
3. It is highly recommended that the cell line used for the assay express the gene of interest (i.e., one in which the regulatory region harboring the variants being tested is functional and the minimal promoter used in the promoter-reporter construct is active). When assaying coding variants in TFs, it is possible to use a cell line that does not express the TF, provided that the TF by itself is capable of activating the promoter of the promoter-reporter construct. This is actually preferable, as the high background activity of an endogenously expressed TF will not interfere with the results. HEK 293 cells are widely used for such assays. However, most TFs work as complexes

and additional constructs that express the essential factors that form the complexes need to be cotransfected. Alternatively, a cell line in which both the TF and the promoter are active can be used for the assay. Note that when studying coding variants based on signal transduction pathways, a cell line in which that pathway is known to exist is preferable.

4. Use of high fidelity polymerase is highly recommended for amplifying the region of interest. Following purification of the cloned constructs, accuracy of the insert sequence must be confirmed by direct sequencing. Use of high fidelity restriction enzymes from NEB is recommended. Most of these enzymes have 100% efficiency in cut-smart buffer offered by NEB and will be helpful for double digestion.
5. The 5' end of the primers used for amplifying the region of interest must contain the selected restriction enzyme sites. It is recommended to include at least four bases 5' of the restriction site on the primers to facilitate digestion. Prior to selecting the restriction enzyme, ensure that the region of interest does not also harbor sites for that enzyme.
6. Treating the double-digested vector with alkaline phosphatase will prevent religation of the linearized vector.
7. The T4 DNA ligase buffer should be vortexed thoroughly before use. This buffer contains ATP, which tends to precipitate.
8. The number of cells to be plated in each well for transfection should be optimized so that they are 60–90% confluent at the time of transfection. Care should be taken to plate equal number of cells in each well, which can be achieved by making a master mix of cells in media and inverting the tube several times and immediately plating the cells into wells. Gently rock the plate back and forth several times to ensure an even spread of the cells.
9. The transfection reagent and constructs should be brought to room temperature before transfection. Mix the transfection reagent by gentle vortexing before making the transfection complex. Transfection complex should be prepared in a serum-free media and transfection should be done under antibiotic-free conditions.
10. The Lipofectamine: DNA ratio needs to be optimized for each cell line and should be determined empirically. In most cases, 3.75 μ L of lipofectamine per 1 μ g of DNA works well when using Lipofectamine 3000.
11. The diluted DNA in the serum-free media should be added into the diluted lipofectamine, and not vice versa. Once the complex is prepared the entire volume should be added

dropwise to the well. Add drops to the media surface and gently rock the plate to ensure even distribution of the transfection complex.

12. Always include a vector-only control and the minimal promoter–reporter construct in the assays. At least two experimental replicates should be included in each plate for each construct. Care should be taken to add equal amounts of each experimental construct to be assayed. The concentration should be determined by repeated measurements.
13. One well containing cells without any treatment should be included to serve as the basal measurement. Equal amounts of the control reporter construct (usually *Renilla* luciferase) should be included in the transfection complex for all wells, except the basal control.
14. It is not necessary to change media after transfection. Cells can be harvested 48 h post-transfection. Before harvesting, look at the cells for any abnormal regions of cell death due to experimental treatment. It is recommended that such wells be omitted from the experiment. While harvesting, ensure complete lysis of the cells.
15. Make $1 \times$ PLB in excess to cover all the wells. Freshly prepared $1 \times$ PLB is recommended. Use of other lysis buffer for DLR assay is not recommended.
16. Prepare LAR II and Stop and Glo reagent in excess. Some samples may need to be assayed multiple times.
17. Vortex all the samples thoroughly before the DLR assay. The length of time of vortexing should be similar for each sample. Clear lysate should be used for the assay. Avoid introducing any cell debris or air bubbles in the LAR II, as these may produce erroneous readings.
18. Mix by pipetting after adding the lysate into LAR II. Do not vortex. The number of times the lysate is mixed by pipetting should be consistent across samples.
19. Experimental conditions must be consistent for all samples, and an equal amount of control reporter must be included in each well. Theoretically, the luciferase activity of the control reporter should be similar for each sample. Large variability in the luciferase activity of the control reporter may indicate decreased transfection efficiency in that well or reduced cell-viability. Such wells should be omitted from the analysis. Any increase or decrease in relative luciferase activity in a sample should be a result of the luciferase activity of the experimental construct, and not a vast drop or increase in luciferase activity of the control reporter. Experimental and biological replicates of that sample should also show a similar effect.

20. Background luminescence can be corrected by measuring the activity of the basal samples. However, most cells have very low background luminescence.

References

- Bush WS, Moore JH (2012) Chapter 11: genome-wide association studies. *PLoS Comput Biol* 8(12):e1002822
- Stranger BE, Stahl EA, Raj T (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 187(2):367–383
- Bick D, Dimmock D (2011) Whole exome and whole genome sequencing. *Curr Opin Pediatr* 23(6):594–600
- Gudbjartsson DF, Helgason H, Gudjonsson SA et al (2015) Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* 47(5):435–444
- Rabbani B, Tekin M, Mahdieh N (2014) The promise of whole-exome sequencing in medical genetics. *J Hum Genet* 59(1):5–15
- Cheung VG, Spielman RS (2009) Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat Rev Genet* 10(9):595–604
- Edwards LS, Beesley J, French JD (2013) Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet* 93(5):779–797
- Westra HJ, Peters MJ, Esko T et al (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* 45(10):1238–1243
- Hindorf LA, MacArthur J, Morales J et al. A catalog of published genome-wide association studies. <http://www.genome.gov/gwastudies>
- Project Consortium ENCODE (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74
- Roadmap Epigenomics Consortium, Kundaje A, Meuleman W et al (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518(7539):317–330
- Brasier AR, Ron D (1992) Luciferase reporter gene assay in mammalian cells. *Methods Enzymol* 216:386–397
- Lorenz WW, McCann RO, Longiaru M, Cormier MJ (1991) Isolation and expression of a cDNA encoding *Renilla reniformis* luciferase. *Proc Natl Acad Sci U S A* 88:4438–4442
- Ow DW, JR WDE, Helinski DR et al (1986) Transient and stable expression of the fire fly luciferase gene in plant cells and transgenic plants. *Science* 234:856–859
- De Wet JR, Wood KV, Deluca M (1987) Fire fly luciferase gene: structure and expression in mammalian cell. *Mol Cell Biol* 7:725–737
- Yun C, Dasgupta R (2014) Luciferase reporter assay in drosophila and mammalian tissue culture cells. *Curr Protoc Chem Biol* 6(1):7–23
- Shifera AS, Hardin JA (2010) Factors modulating expression of *Renilla luciferase* from control plasmids used in luciferase reporter gene assays. *Anal Biochem* 396(2):167–172
- GTEX Consortium (2015) Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348(6235):648–660
- Ward LD, Kellis M (2016) HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res* 44(D1):D877–D881
- Boyle AP, Hong EL, Hariharan M et al (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 22(9):1790–1797
- Felekkis K, Touvana E, Stefanou C (2010) microRNAs: a newly described class of encoded molecules that play a role in health and disease. *Hippokratia* 14(4):236–240
- SL H, Cui GL, Huang J (2016) An APOC3 3'UTR variant associated with plasma triglycerides levels and coronary heart disease by creating a functional miR-4271 binding site. *Sci Rep* 6:32700

Part IV

Identification of Disease Genes

Physiologic Interpretation of GWAS Signals for Type 2 Diabetes

Richard M. Watanabe

Abstract

This chapter reviews both statistical and physiologic issues related to the pathophysiologic effects of genetic variation in the context of type 2 diabetes. The goal is to review current methodologies used to analyze disease-related quantitative traits for those who do not have extensive quantitative and physiologic background, as an attempt to bridge that gap. We leverage mathematical modeling to illustrate the strengths and weaknesses of different approaches and attempt to reinforce with real data analysis. Topics reviewed include phenotype selection, phenotype specificity, multiple variant analysis via the genetic risk score, and consideration of multiple disease-related phenotypes. Type 2 diabetes is used as the example, not only because of the extensive existing knowledge at the genetic, physiologic, clinical, and epidemiologic levels, but also because type 2 diabetes has been at the forefront of complex disease genetics, with many examples to draw from.

Key words Quantitative traits, Genome-wide association, Mathematical modeling, Genetic risk score, Statistics, Regression analysis, Phenotyping

1 Introduction

Novo Nordisk began an advertising campaign in 2009 built around the phrase “There may be 2 types of diabetes, but there’s more than 2 types of patients with diabetes.” This very simple message highlighted the challenge faced by clinicians struggling to treat and prevent diabetes, as each individual patient presents with different challenges in dealing with their diabetes or trajectory toward diabetes. Although categorized into two broad groups, in reality, diabetes is a group of diseases that manifest in fasting hyperglycemia. In particular, type 2 diabetes involves a complex interplay of genetic risk, lifestyle, and socioeconomic factors that have been literally investigated for centuries. It is easy to forget that the first description of diabetes as a medical condition was as early as 1500 B.C [1] and research on diabetes has stretched across hundreds of years. Sugar in the urine of patients with diabetes had been suspected for many years, but it was not until 1674 that

Charles Willis actually noted the urine of his patients with diabetes had a sweet taste [1]. Additional milestones in diabetes research include contributions of individuals like Claude Bernard (1813–1878), who identified the liver as an integral organ involved in diabetes, Paul Langerhans (1847–1888), who described the structure of the pancreatic islets that bear his name, Oskar Minkowski (1858–1931) and Josef von Mering (1849–1908), who discovered the critical importance of the pancreas in diabetes, Sir Frederick Banting (1891–1941) and Charles Best (1899–1978), who discovered insulin, Frederick Sanger (1918–2013), who determined the amino acid sequence of insulin, Rosalyn Yalow (1921–2011) and Solomon Berson (1918–1972), who developed the radioimmunoassay to measure insulin, and countless other individuals who contributed to our current knowledge of the pathophysiology of diabetes.

Much of the knowledge gained regarding type 2 diabetes has been translated to relatively effective treatments for the disease, but success has not been uniform. Physicians grapple with countless patients who are difficult to treat using recommended protocols. Furthermore, much of the knowledge has not translated to effective disease prevention, which is exemplified by the continuing rise in incidence and prevalence of the disease [2, 3]. Numerous studies under highly controlled conditions have shown that lifestyle and pharmacologic interventions can be effective in significantly reducing risk for future diabetes in at-risk individuals [4–8]; however, for a variety of reasons many of these approaches have been difficult to translate into clinical practice and even less so into the community.

In 1962, Dr. James Neel proposed the “thrifty gene hypothesis” to explain the prevalence, heterogeneity, and severity of diabetes [9]. Neel would eventually call diabetes “the geneticist’s nightmare” a phrase that captured the complex nature of the disease: polygenic with low genetic effects and complex interactions with environment [10]. The field of human genetics has changed considerably since the coining of that phrase. The successful application of linkage analysis to identify genetic loci contributing to risk for monogenic disease led to its application to complex diseases, albeit with relatively little success. The lack of success was partly explained by Risch and Merikangas, who formally showed genetic association was statistically more powerful than linkage analysis [11]. However, technical limitations and significant knowledge gaps regarding the human genome made implementation of genome-wide association (GWA) more of a dream than a possibility. But rapid improvements in genotyping and computing technology, increased knowledge about the human genome through the Human Genome [12, 13] and HapMap Projects [14] plus other advances made GWA a reality, leading to a “golden age” for complex disease genetics. Numerous GWA studies have identified thousands of genetic variants across the genome showing strong

evidence of contributing to risk for disease or variation in disease-related quantitative phenotypes. The rapid growth in the number of such loci led to the establishment of the GWAS Catalog (<http://www.ebi.ac.uk/gwas/>), a repository of results from published GWA studies.

In many ways, the study of the genetics of type 2 diabetes was at the forefront of this genetic revolution. As of this writing, nearly 100 loci have been identified as contributing to risk for type 2 diabetes [15–26] and hundreds of loci have been identified as contributing to variation type 2 diabetes related quantitative phenotypes [27–48]. The genomic chips created for GWA analysis were designed to capture common variation, typically minor allele frequency (MAF) greater than 5%, which is a limitation of GWA studies. This led to the phenomenon of “missing heritability” [49]; the identified genetic variation does not account for all of the heritable variation. It was thought that next generation sequencing of whole human exomes and genomes would identify low frequency variants with larger genetic effect size that would account for this “missing heritability”, but initial results from these studies indicate the effects of rare variants may not be substantially greater than that of higher frequency variants [50]. Regardless, the GWA and next generation sequencing studies have not just illuminated new biologic pathways contributing to the pathogenesis of diabetes, but have also reinforced and provided additional insights into previously known pathways and revealed new potential pharmacologic targets.

There are many paths to take once a variant is shown to be associated with type 2 diabetes or type 2 diabetes-related phenotypes. One approach is to examine the role of these variants in the pathophysiology of the disease, which is the focus of this chapter. In particular, we review approaches to understanding how genetic variation might alter human physiology to contribute to the pathogenesis of type 2 diabetes. The complex relationship between genetic variation and physiology, so-called genotype–phenotype relationships, is critical in the determination of how genetic variation may be leveraged to improve interventional strategies, both pharmacologic and lifestyle, and reduce diabetes-attributable morbidity and mortality.

2 Methods

We will focus on three general issues in the assessment of the role of genetic variation in the pathogenesis of type 2 diabetes: (1) consideration of phenotypes, (2) phenotype specificity, and (3) the utilization of the genotype risk score. Specific issues surrounding each will be discussed with a mathematical model used for illustrative

purposes. The results from computer simulations performed using the mathematical model are further reinforced by practical examples derived from analysis of data from the BetaGene Study [51].

2.1 The BetaGene Study

BetaGene is a study of Mexican American families of probands with and without a prior diagnosis of gestational diabetes mellitus (GDM). The broad sampling design for this study was based on the fact that Mexican American women with prior GDM have a 55% five-year post-partum risk of developing type 2 diabetes [52] with similar levels of elevated risk observed in other populations [53]. BetaGene provides a unique opportunity to study the relationship between genetic variation and type 2 diabetes-related phenotypes in the context of elevated disease risk. The framework of sampling based on previous GDM status ensured representation of individuals with and without elevated risk for development of type 2 diabetes, i.e., a wide spectrum of glucose tolerance would be observed. It should be noted that the BetaGene sample specifically excludes individuals with type 1 or type 2 diabetes and families in which individuals presented as being GAD antibody positive.

Another unique characteristic of the BetaGene Study was the performance of detailed phenotyping, which provides quantitative measures of type 2 diabetes-related phenotypes not found in most human genetic studies. These include body composition by dual-energy X-ray absorptiometry (DXA), oral glucose tolerance test (OGTT) with blood samples obtained every 30-min post-ingestion for 2 h, and frequently sampled intravenous glucose tolerance tests (FSIGT) analyzed by the Minimal Model [54] to obtain quantitative measures of glucose effectiveness, insulin sensitivity, insulin secretion, and pancreatic beta-cell function. These provide direct quantitative measures of key phenotypes known to contribute to the pathogenesis of type 2 diabetes, in addition to traditional clinical phenotypes.

2.2 Leveraging Mathematical Modeling

We provide additional illustration of issues related to physiologic analysis of genetic variation by leveraging mathematical modeling. We previously introduced a model describing the gluco-regulatory system [55]. The model links existing validated models of glucose and insulin kinetics to create a single unified model that accurately reflects the closed-loop feedback relationship between glucose and insulin (Fig. 1). Because each parameter in the model has direct physiologic interpretation and an underlying population distribution, we can model the effect of genetic variation by defining a genotype-specific parameter distribution underlying the overall population distribution. The magnitude of the genetic effect for a given variant can be modeled as a function of the differences in the genotypic means.

This modeling framework allows us to do two things. First, the model is designed to allow simulation of any clinical phenotyping

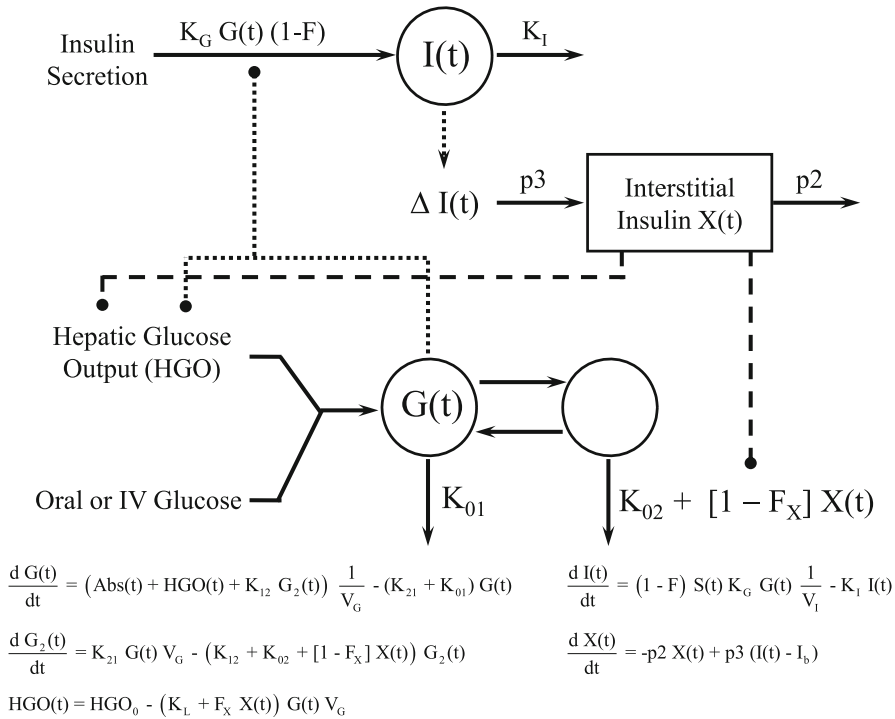


Fig. 1 Model of Gluco-regulation. Structure of the model along with the system of differential equations is shown. Compartment G represents the concentration of glucose in plasma and tissues that rapidly equilibrate with plasma (mg/dL, insulin insensitive), while G_2 represents the mass of glucose in tissues that slowly equilibrate with plasma (mg, insulin sensitive). V_G is the volume of the extracellular space (glucose distribution volume). $\text{Abs}(t)$ is the rate of absorption of glucose from the gastrointestinal tract and $\text{HGO}(t)$ is the endogenous glucose production rate. K parameters are fractional transfer coefficients (min^{-1}). F_X represents the fraction of the effect of interstitial insulin that acts to suppress endogenous glucose output, while $(1 - F_X)$ is the remaining fraction which accelerates glucose disposal from the peripheral compartment. The rate of endogenous glucose production is limited by plasma glucose ($G(t)$) and interstitial insulin ($X(t)$). HGO_0 is the endogenous glucose output in the absence of effects of interstitial insulin and/or plasma glucose to restrain glucose production. $S(t)$ is the moment-by-moment secretion rate of insulin from the pancreatic islets; F_I is the fractional clearance rate of secreted insulin by the liver. V_I is the distribution volume of insulin in the body. I_b is the basal (fasting) insulin concentration. The effect of insulin on glucose kinetics was modeled as a remote insulin effect, and partitioned into two components (F_X and $1 - F_X$). Pre-hepatic insulin secretion is a known input to this model. The insulin secretion rate changes in proportion to the ambient glucose concentration, with K_β representing β -cell sensitivity to glucose. The term $(1 - F)$ describes the fraction of insulin secretion that survives hepatic transit, with F_I representing fractional hepatic insulin extraction. The plasma insulin concentration above basal determined by this model is used to determine the effect of interstitial insulin, $X(t)$

protocol used in modern type 2 diabetes research. This includes clinical protocols such as the OGTT, FSIGT, or euglycemic glucose clamp. Second, because the characteristics of the population distribution for each model parameter is known and the genotype-specific distributions underlying the population can be created, it is possible to draw genotype-specific samples from the distribution.

These characteristics allow us to draw samples from the population that have known characteristics and allow us to “phenotype” these samples under highly controlled conditions. The simulation framework allows us to assess various aspects of the contribution of genetic variation to the pathophysiology of type 2 diabetes and assess the complexities underlying genotype–phenotype relationships.

3 Results

3.1 *Lack of Overlap in Genetic Loci*

One of the strengths of studying the genetics of type 2 diabetes is the long history of clinical and epidemiologic studies performed to date and the extensive array of type 2 diabetes-related quantitative phenotypes that have been measured. Furthermore, it is well established that in most populations, type 2 diabetes is characterized by relative obesity, insulin resistance, and pancreatic β -cell dysfunction. Thus, measures of these phenotypes, including more traditional clinical phenotypes, provide opportunities to leverage quantitative trait analysis to better understand the genetics of the disease. In fact, the first type 2 diabetes risk locus to be identified by primary analysis of a diabetes-related phenotype was melatonin receptor-1B (*MTNR1B*), which was initially shown to be associated with fasting glucose levels, and later with risk for type 2 diabetes [37, 40]. This demonstrated that type 2 diabetes risk loci could be identified by analyzing type 2 diabetes-related phenotypes, with the added bonus of being able to make improved physiologic inference, given the knowledge that the locus was associated with both disease and a specific disease-related trait.

However, not all disease-related quantitative trait loci contribute to diabetes risk and not all diabetes risk loci contribute to variation in known type 2 diabetes-related quantitative traits. The former is not surprising, since one would expect that a GWA study of a diabetes-related quantitative trait might identify loci that only contributed to the variation in that trait and not necessarily contribute to disease risk. However, the latter is a bit more puzzling, as one might assume that a diabetes risk locus should also be associated with a known type 2 diabetes-related quantitative trait. However, GWA of disease is an agnostic approach that makes no a priori assumption regarding the underlying biology other than the presence or absence of disease, and therefore could identify loci underlying phenotypes not yet examined in the context of type 2 diabetes. In fact, when one compares diabetes risk loci with diabetes-related quantitative trait loci, there is surprisingly little overlap. Grarup and colleagues examined the overlap in identified loci across type 2 diabetes, BMI, waist circumference, waist-to-hip ratio, fasting glucose, and fasting insulin [56] and found very little overlap in loci. Among the 45 loci associated with variation in

fasting glucose, only 18 also contributed to risk for type 2 diabetes. Likewise, although obesity is a significant risk factor for type 2 diabetes, only 4 of 40 loci associated with this trait also contributed to risk for type 2 diabetes.

One explanation for this lack of overlap is that the disease risk locus may be associated with a phenotype that is not among the traditional clinical phenotypes measured in the majority of studies of diabetes. This is actually one of the weaknesses of genetic studies of disease, in that decisions regarding phenotypic measurements are typically driven by clinical knowledge and thus, clinically relevant phenotypes tend to be measured. Alternatively, we might leverage archival samples and therefore be constrained by the protocols employed and the samples archived. Furthermore, a locus may be revealed in stimulated states, as opposed to the fasting state. In which case, unless the appropriate clinical protocol has been employed, the investigator will again be constrained.

3.2 Modeling the Effect of Peroxisome Proliferator-Activated Receptor- γ (*PPARG*)

We turn to computer simulation to illustrate how a nontraditional phenotype might be optimal for assessing association with a disease risk locus. The Pro12Ala variant (rs1801282) in *PPARG* was one of the first type 2 diabetes risk loci to be identified [57, 58]. However, for many years, evidence for its association with diabetes-related phenotypes was equivocal. While this was partly attributed to the lack of statistical power stemming from relatively small sample sizes, the lack of assessing appropriate phenotypes may have also contributed to the inconsistent results. We illustrate this by simulating the effect of *PPARG* Pro12Ala using our computer model. We assume a genetic variant underlying parameter p_3 , which is involved in insulin sensitivity, with 15% MAF, and a genetic effect of 30% of the maximum difference between homozygous genotypes. Genotype-based p_3 values were randomly selected from the population distribution of p_3 and OGTTs were simulated employing the selected variant. The process was repeated multiple times to simulate a sample drawn from a population. Three thousand replicates of 1000 OGTTs were simulated and the power to detect association between individual OGTT glucose values and the simulated genetic variant was assessed. All other model parameters were fixed at their population averages to ensure the only source of genetic variation was from the simulated *PPARG* variant and random Gaussian noise was added to the simulated glucose data to simulate assay variation. The simulated glucose data was tested for association with the simulated genetic variant by linear regression and statistical power to detect association was assessed.

The standard clinical protocol for the OGTT calls for samples to be taken at fasting and 2 h post-load. Thus, most studies have glucose, and sometimes insulin, measured at these two time points. However, computer simulation allows us to simulate glucose values for any desired time-point during a 2-h OGTT. Thus, we

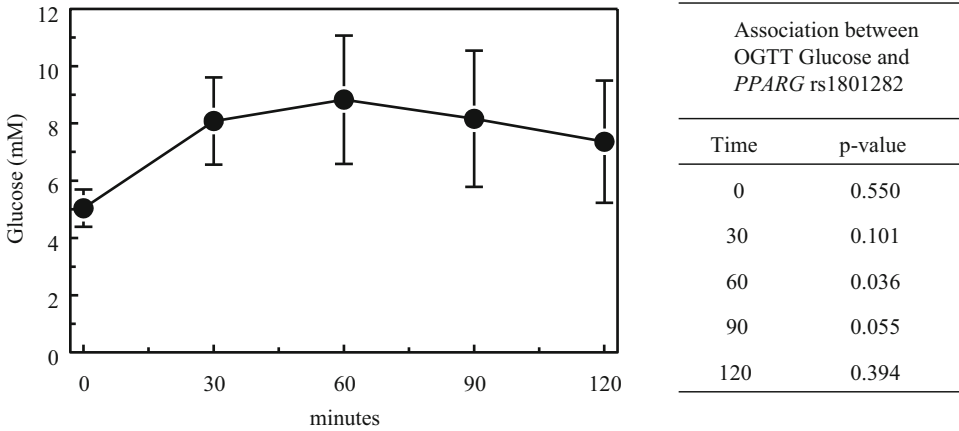


Fig. 2 Average oral glucose tolerance test glucose values (mean ± SD) for 1630 individuals from the BetaGene study. Table on the right shows the *p*-value for the test of association between *PPARG* rs1801282 and each OGTT glucose time point adjusting for age, sex, and percentage body fat

specifically compare the power to detect association between the simulated 1- and 2-h glucose values, the former rarely measured in most studies. The simulation results show that the 1-h glucose value had greater statistical power to detect association with the simulated genetic variant compared to the 2-h glucose value (0.83 vs. 0.74). This power analysis suggests that if one were to select a phenotype optimized for detecting association with *PPARG* rs1801282, one would have been best served by collecting 1-h glucose samples from the OGTTs.

We next attempted to replicate the simulation findings by analyzing data from the BetaGene study. Blood samples were collected every 30 min post-load in the OGTTs performed in the BetaGene study, unlike most studies that only perform the 2-point OGTT. Fig. 2 shows the average glucose profile on the left (mean ± SD) and the corresponding *p*-value for the test of association between the different glucose time points and rs1801282. The test for association was performed adjusting for age, sex, and percentage body fat. Despite the relatively small sample size ($n = 1630$), the analysis of BetaGene data clearly shows the 1-hour glucose level having the strongest evidence for association with rs1801282 ($p = 0.036$) compared to the 2-h ($p = 0.394$) or other glucose values. In fact, even the 90-min glucose value showed better evidence for association ($p = 0.055$) compared to the fasting or 2-h values. This simple example shows that atypical phenotypes may be associated with type 2 diabetes risk loci and may partly explain the lack of overlap with GWA studies of type 2 diabetes-related phenotypes. It also suggests that careful consideration must be made in phenotype selection when studying genetics. The standard set of clinically based phenotypes may not be

sufficient to untangle the contribution of genetic variation to the pathogenesis of type 2 diabetes and identify the optimal targets of interventional development.

3.3 Phenotype Specificity

The issue of analyses being limited to clinically relevant phenotypes raises important questions regarding appropriate and specific phenotypes. As noted, insulin resistance and beta-cell dysfunction are hallmarks of type 2 diabetes. However, accurately assessing these phenotypes requires clinical protocols beyond simple fasting blood draws or OGTTs. For example, the euglycemic glucose clamp is the gold standard method for quantifying insulin resistance. However, the cost and complexity of this protocol makes it less favorable for large-scale human studies. This has led many to use indirect measures of these phenotypes, as exemplified in reports from the MAGIC consortium [59, 60]. However, the utility of these indirect measures is controversial [55, 61–63], and even if highly correlated with direct measures of these phenotypes, there are issues as to whether they capture the same genetic information [64, 65]. The use of indirect phenotypes can lead to issues in physiologic interpretation.

We present two examples of how indirect phenotypes can lead to difficulties in physiologic interpretation. If a genetic variant has implications for a given phenotype, say insulin resistance, then one would expect that variant to show association with any phenotype that directly reflects insulin resistance. The MAGIC consortium examined the association between 37 type 2 diabetes risk loci and multiple type 2 diabetes-related phenotypes to assess the physiologic implications of genetic variation contributing to type 2 diabetes risk [60]. Insulin resistance was a trait of primary interest, but direct measures of insulin resistance/sensitivity were only available in a small subset of participating studies and among those, the methods of measurement varied and included FSIGT with Minimal Model [54, 66], islet suppression test [67], and euglycemic glucose clamp [68]. Five indirect measures of insulin resistance were computed across all studies based on fasting and/or OGTT data; HOMA-IR [69], the Stumvoll Index [70], the Belfiore Index [71], the Matsuda Index [72], and the Gutt Index [73] to maximize statistical power for analysis of insulin resistance as a phenotype. The three direct measures were examined as a single group, thus there were six different “measures” of insulin resistance/sensitivity examined. Marker rs7578326 in *IRSI* was the only locus that showed evidence for association with all six measures of insulin resistance [60]. Other loci, such as rs13081389 in *PPARG*, showed evidence for association with only a subset of these traits. The heterogeneity in outcomes for a given SNP raises an important question. If all six measures are supposed to reflect insulin resistance, why are the association results not consistent? In the case of the indirect measures, statistical power is not likely to explain the

heterogeneity, because the sample sizes were large ($>10,000$ samples) and relatively similar across phenotypes. If any single phenotype were to be inconsistent, it would have been the combined direct measure group, since the sample size was approximately 40% of that for the indirect measures and therefore would have suffered from significantly low statistical power. One reason could be confounding effects from other phenotypes, such as insulin secretion, which we previously showed was a significant confounder for many of these indirect measures [55]. Alternatively, these indirect measures, while phenotypically correlated with direct measures of insulin resistance, may not capture the same genetic information. For example, two studies have examined the phenotypic, genetic, and environmental correlation between HOMA-IR and the euglycemic glucose clamp [64] and the FSIGT with Minimal Model analysis [65]. In both cases, the phenotypic correlation between HOMA-IR and the direct measure was high, but the genetic correlation was significantly lower [65, 64]. This suggests that while HOMA-IR may be phenotypically correlated with the euglycemic clamp or Minimal Model-based S_I and useful as an overall alternative measure of insulin resistance, it may not capture the same genetic information and therefore may not be as useful in genetic studies.

The second example takes advantage of our mathematical model to illustrate how the confounding effects of other phenotypes, such as insulin secretion [55], can confound the interpretation of SNP association results when examining insulin resistance. We simulated the effect of two different SNPs on the glucoregulatory system; one that affects insulin sensitivity and another that affects pancreatic beta-cell function. We simulated the effect of *PPARG* rs1801282, as described above, as a variant affecting insulin sensitivity. We simultaneously simulated the effect of rs10830963 in *MTNR1B* on pancreatic beta-cell function via parameter K_G (cf. Fig. 1). We assumed a 20% MAF and a genetic effect 50% of the maximum for the simulated *MTNR1B* variant. Paired OGTTs and FSIGTs were simulated based on randomly drawn pairs of p_3 and K_G for the two simulated SNPs. As before, all other model parameters were kept constant and glucose and insulin data were simulated for the two clinical tests. Random Gaussian error was added to the simulated glucose and insulin data to simulate assay variation. The Stumvoll Index of insulin sensitivity was computed from the simulated OGTT data [70] and the simulated FSIGT data were analyzed using the Bergman Minimal Model [54, 66] to estimate S_I . The simulated SNP genotype data were then tested for association with the two different insulin sensitivity indices. One hundred replicates of 1000 or 2500 OGTT/FSIGT pairs were examined to assess power to detect association.

The median correlation between the Stumvoll index computed from the simulated OGTT and S_I computed from the simulated

FSIGT was 0.70 (range: 0.61–0.76), which is similar to values reported in the literature. Results based on the 1000 replicates were similar to those obtained for 2500 replicates, so only results based on the 2500 replicates are described. Power to detect association between the simulated *PPARG* rs1801282 was 37% for the Stumvoll Index and 47% for S_I . The moderately higher statistical power with S_I likely reflects the fact that S_I is a direct measure of insulin sensitivity. Because both the Stumvoll Index and S_I are measures of insulin sensitivity, one would expect both indices to show weak power to detect association with genetic variants whose primary effect is on phenotypes unrelated to insulin sensitivity. However, when examining the association between the simulated *MTNR1B* rs10830963 and the two indices, the Stumvoll Index had substantially greater power to detect association compared to S_I (43% vs. 21%). In fact, power to detect association with the simulated beta-cell variant (*MTNR1B*) was similar to, if not slightly greater than, the power to detect association with the insulin sensitivity variant (*PPARG*; 43% vs. 37%). In contrast, S_I from the Minimal Model had substantially lower power to detect association with the simulated beta-cell variant compared to the insulin sensitivity variant (47% vs. 21%), which is better aligned with what one might expect from a phenotype reflecting insulin sensitivity.

We compared our simulation results with results from the BetaGene study. We tested association between *PPARG* rs1801282 and *MTNR1B* rs10830963 with the Stumvoll Index and Minimal Model S_I in 1093 BetaGene participants with complete data. The test for association included adjustment for age, sex, and percentage body fat. The correlation between the Stumvoll Index and S_I in BetaGene was 0.58, somewhat lower than what was observed in the simulation study. The observed MAF for *PPARG* rs1801282 was 0.11 and for *MTNR1B* rs10830963 was 0.22. The Stumvoll Index and S_I showed no evidence for association with *PPARG* rs1801282 ($p = 0.16$ and $p = 0.66$, respectively). This outcome is not surprising, given the relatively small sample size and the presumed low genetic effect size of *PPARG*. However, both indices are consistent in outcome and consistent with our simulation results. However, the Stumvoll Index showed evidence for association with *MTNR1B* rs10830963 ($p = 0.043$), while S_I did not ($p = 0.86$). This result is consistent with our simulation results, which suggested the Stumvoll Index had near-equivalent power to detect association with genetic variants affecting insulin sensitivity or pancreatic beta-cell function.

The simulation and real data analysis emphasize the importance of phenotype specificity, especially with respect to interpretation of the results. In the case of BetaGene, had one only performed analyses using the Stumvoll Index, one might draw the conclusion that *MTNR1B* rs108309063 is involved in regulation of insulin sensitivity, when there is sufficient evidence suggesting otherwise

[39, 74–77]. Specificity in phenotypes is critical in dissecting the physiologic implication of genetic variation. Thus, additional investigation is necessary to understand the genetic relationship between indirect and direct measures of various disease-related phenotypes. A simple overall phenotypic correlation is likely insufficient and more detailed genetic correlation must be assessed to utilize certain phenotypes with confidence.

3.4 The Genotype Risk Score

The discovery of multiple variants associated with risk for type 2 diabetes has resulted in a need for simple approaches to assessing the association between phenotypes and the net effect of genetic variation. The genotype risk score (GRS) has become a widely used approach to assess the contribution of known genetic risk loci in genetic association studies. The concept and assumptions behind the GRS are simple. Each risk allele at a specific locus makes an additive contribution to the overall genetic risk. Therefore, counting the total number of risk alleles across all risk loci should provide an estimate of the overall genetic burden carried by a given individual. The GRS can be unweighted, simply summed up, or weighted, weighting the number of alleles by an estimate of their effect sizes. While the latter is the most accurate approach to use, many times independent estimates of individual locus effect sizes are not available, leading to the wide-spread use of the unweighted GRS.

While the GRS is a simple and effective tool to assess overall genetic effects when examining type 2 diabetes, many investigators have used the same GRS to test for association with type 2 diabetes-related phenotypes. This can be problematic, as there is no guarantee the summed effect of risk loci will be associated with a given disease-related phenotype. This is partly exemplified by the lack of overlap in loci between disease and disease-related traits. Additionally, while the direction of effect is consistent with respect to disease risk when constructing a GRS based on type 2 diabetes risk loci; all loci increase risk for type 2 diabetes, there is no guarantee the direction of effect will be consistent when examining type 2 - diabetes-related traits. Thus, instead of enhancing power, one could be creating undue heterogeneity that might significantly reduce power.

We demonstrate these issues by computing an unweighted GRS using 56 type 2 diabetes risk loci identified from the literature and genotyped in the BetaGene study. The GRS was tested for association with type 2 diabetes-related phenotypes adjusting for age and sex. We also performed stepwise regression analysis using forward-backward selection to identify a parsimonious model that included the “best” subset of SNPs accounting for variation in a given type 2 diabetes-related phenotype. This contrast can be broadly viewed as a comparison between a regression model that includes all 56 SNPs versus the parsimonious model that includes only the “best” subset of SNPs.

Table 1
Comparison of quantitative trait association results based on GRS vs. Stepwise regression

Trait	GRS ^a		Stepwise regression		
	% variation ^b	<i>p</i> -value	# Of SNPs ^c	% variation ^b	<i>p</i> -value
Body mass index	0.16%	0.219	12	4.4%	1.4×10^{-5}
Body fat percent	0.12%	0.141	13	2.8%	2.3×10^{-6}
Waist hip ratio	0.01%	0.686	10	3.2%	2.5×10^{-5}
Fasting glucose	0.87%	0.004	11	5.1%	8.0×10^{-7}
2-Hr. glucose	0.42%	0.039	9	4.2%	1.4×10^{-6}
Fasting insulin	0.03%	0.605	15	8.2%	5.1×10^{-11}
30-Minute insulin	1.7%	6.0×10^{-5}	17	10.0%	1.8×10^{-14}
2-Hr. insulin	0.03%	0.622	17	6.5%	9.7×10^{-8}
Glucose effectiveness	1.00%	0.002	16	6.4%	6.9×10^{-8}
Insulin sensitivity	0.03%	0.585	11	5.0%	9.9×10^{-7}
Acute insulin response	3.0%	5.4×10^{-8}	17	12.8%	$<1 \times 10^{-16}$
Disposition index	2.9%	6.7×10^{-8}	15	9.6%	5.6×10^{-15}
Cholesterol	0.06%	0.438	11	5.2%	7.9×10^{-8}
HDL cholesterol	0.00%	0.949	13	5.9%	2.1×10^{-8}
LDL cholesterol	0.17%	0.189	11	4.5%	2.1×10^{-6}
Triglycerides	0.02%	0.667	8	3.1%	3.1×10^{-5}
Systolic blood pressure	0.47%	0.022	11	4.1%	1.5×10^{-6}
Diastolic blood pressure	0.09%	0.325	15	4.8%	3.5×10^{-6}

^a GRS based on 56 known type 2 diabetes risk loci

^b Proportion of the total trait variation explained

^c The total number of SNPs in the final regression model

Table 1 summarizes the results of the analysis and clearly reveals the GRS does not show evidence for association with most type 2 diabetes-related phenotypes, except for fasting glucose, 2-h glucose, 30-min insulin, glucose effectiveness, acute insulin response, disposition index, and systolic blood pressure. Broadly, the GRS accounted for a relatively small proportion of trait variation (0.02–3%), with the GRS maximally accounting for 3% of the variation in acute insulin response. If one were to attempt a physiologic interpretation of the GRS-based association results, one might conclude that type 2 diabetes loci affect insulin secretion and pancreatic beta-cell function, which in turn alters glucose levels. The systolic blood pressure association is harder to include in a broader physiologic picture. This interpretation is not without

merit, since the majority of type 2 diabetes risk loci map to the pancreatic beta-cell [78, 79]. However, clinical and epidemiologic studies clearly show a significant contribution of obesity and insulin resistance to the pathogenesis of type 2 diabetes. Yet, the GRS shows no evidence for association with those traits in our analysis.

Stepwise regression analysis paints a different picture. The individual stepwise regression models identifying the “best” subset of SNPs associated with individual diabetes-related traits include models with 8–17 of the 56 SNPs accounting for 2.8–12.8% of the trait variation (Table 1). The increase in the proportion of variation explained is partly due to the fact that stepwise regression is designed to identify the subset of SNPs that maximizes this proportion. However, these results also highlight that a significant proportion of the variation in individual type 2 diabetes-related traits can be accounted for given the right set of SNPs.

Also in contrast to the GRS approach, stepwise regression results for each individual trait showed overall evidence for association, with p-values significantly smaller than those observed for the GRS. We now start to see associations with obesity, insulin resistance, and other type 2 diabetes-related phenotypes with stepwise regression. Physiologic interpretation of these results is much more complex, since different sets of SNPs show evidence for association with different type 2 diabetes-related phenotypes. In fact, different sets of SNPs appear to be associated with what may appear to be physiologically similar phenotypes. Fig. 3 shows an example where the list of SNPs in the final regression model for fasting insulin, OGTT 30-min insulin, OGTT 2-h insulin, and acute insulin response from the FSGIT are compared. The a priori expectation might be that there should be considerable overlap among SNPs across the various models, since all four traits are insulin-based and three are from the insulin response from the same clinical test, the OGTT. However, only one SNP, rs7754840 in *CDKALI*, appears in all four models. Similarly, there is only one SNP that overlaps between 30-min OGTT insulin and 2-h OGTT insulin (rs12454712 in *BCL2*). Similar comparisons can be made across different sets of traits with similar results. The stepwise regression analysis reveals that type 2 diabetes risk variants underlie a variety of different type 2 diabetes-related phenotypes and that the proportion of variability in these traits accounted for by these SNPs can be substantial.

Another important observation from the stepwise regression analysis is that only 53 of the 56 SNPs were included in a final model of any of the traits; rs1084299 in *SLC16A12*, rs391300 in *SRR*, and rs791595 in *LOC105375494* showed no evidence of association with any of the traits examined. This suggests these three SNPs may be associated with phenotypes not examined in this analysis. Also, while the majority of SNPs were included in final models for multiple traits, three SNPs stand out as being included

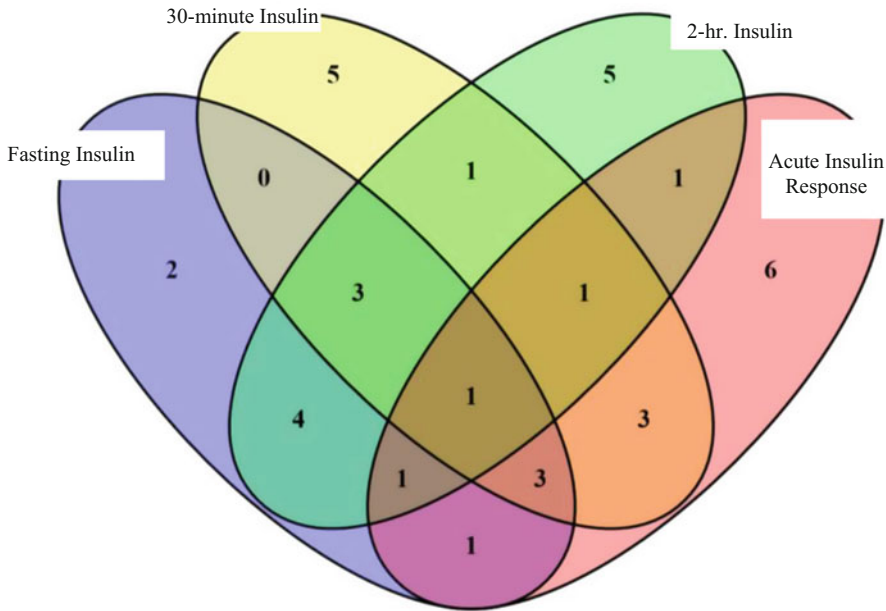


Fig. 3 Venn Diagram showing the overlap in SNPs selected by stepwise regression for traits involving insulin. Final stepwise regression models for four different insulin traits, fasting insulin, 30-min OGTT insulin, 2-h OGTT insulin, and acute insulin response, were compared. Overlap is minimal across all four traits and only rs7754840 in *CDKAL1* appeared in regression models for all four traits

in final models for >10 of the 18 traits examined; rs10830963 in *MTNR1B* was associated with 11 traits, rs11708067 in *ADCY5* was associated with 11 traits, and rs243021 upstream of *BCL11A* was associated with 12 traits. This suggests these variants may have broad physiologic effects and may be primary targets for interventional development.

4 Final Thoughts

Type 2 diabetes was known to be difficult to clinically manage and even harder to prevent even in the absence of genetic knowledge. The recent gain in genetic knowledge promises to bring us closer to the concept of personalized medicine. While individualizing medical care might be a stretch, the idea of personalizing treatment and prevention to subgroups within the population is an achievable goal. Improvements in the diagnosis and treatment of rare forms of diabetes, such as neonatal diabetes [80, 81], have laid the foundations for application of genomics to the treatment and prevention of type 2 diabetes.

The road forward will be challenging. Despite the long history of studying type 2 diabetes and the recent gains in genetic knowledge, the disease still remains the geneticist's nightmare. The

traditional reductionist approach to fine-mapping loci identified by recent GWA and whole genome sequencing studies, followed by functional characterization of associated variants, continues to be an important and necessary component to understanding the contribution of genetic variation to disease risk. However, understanding the broader physiologic implications of genetic variation remains equally important. It is easy to forget that complex diseases such as type 2 diabetes involve multiple systems working in complex ways to tightly regulate glucose, and that small changes in one part of the system can easily be compensated for by changes in other parts of the system. Thus, understanding the complex interplay among these systems and how genetic variation alters them, will be paramount to identifying successful treatment or prevention regimens. In other words, continued study of *in vivo* physiology in human subjects will continue to be a critical component of the research portfolio moving forward.

We have shown that the overlap of genetic loci among related phenotypes is not large, allowing us to draw several conclusions. First, there are clearly important phenotypes that have not been identified that could provide important insights into the disease. We showed by both computer simulation and real data analysis that examining nontraditional phenotypes can provide important insights into the underlying physiology of disease. Of course, partly paraphrasing former Secretary of Defense Donald Rumsfeld, these phenotypes form part of the “unknown unknowns—the ones we don’t know we don’t know” and identifying these new biomarkers will be one of the great challenges of ongoing diabetes research.

Second, part of our inability to advance our knowledge of the genetic underpinnings of physiology is the persistent use of inferior phenotypes that can mislead our interpretation of results. Again, we used computer simulation and real data analysis to illustrate this problem and we hope that we have convinced the reader of the need to carefully consider the quality of the phenotype being used and how any association result fits into the larger puzzle. As an additional example of the importance of phenotypes, we noted above that type 2 diabetes risk alleles in *MTNR1B* were first identified through the analysis of fasting glucose [37, 40]. These associations were identified through the analysis of tens of thousands of samples [37, 40], and subsequent studies resulted in the conclusion that *MTNR1B* variants exert their effects at the level of the pancreatic beta-cell [39]. Thus, it is currently believed that *MTNR1B* variants alter insulin secretion or pancreatic beta-cell function, which subsequently alters glycemic levels, leading to hyperglycemia. However, a subsequent GWA study examining the acute insulin response as a phenotype was able to convincingly show evidence for association between variation in *MTNR1B* and acute insulin response and disposition index with approximately 2500 samples [82]. This study demonstrates the utility of refined and specific

phenotypes for genetic studies, i.e., greater bang for the buck. Thus, we need to carefully consider phenotyping in future studies, if we are to fully dissect the pathophysiology of type 2 diabetes.

Additional studies will be required to better understand how these various loci fit into the larger puzzle of the pathophysiology of the disease. We showed that simplistic methods, such as the GRS, are not likely to be helpful in quantitative trait analysis, which is where we need to concentrate if we are to elucidate the role of genetic variation. It should be noted that it is appropriate if a GRS is created using quantitative trait loci and used to analyze quantitative traits, but creation of a GRS based on disease risk loci for the purpose of analyzing disease-related quantitative traits is problematic, as shown here. We contrasted GRS with stepwise regression, which is a classically useful approach to identify key predictors of specific phenotypes. The stepwise analysis revealed an additional level of complexity in that there is little overlap in loci affecting multiple phenotypes. We showed that among the loci examined, three appear to be important to multiple phenotypes, suggesting these loci may be critical nodes in a wider network of interactions among loci to alter the trajectory to disease. Additional analyses will be needed to advance this framework.

The last point brings forth two topics that were not touched upon in this chapter, but are important approaches to consider: gene–gene and gene–environment interactions. These interactions cannot be ignored, given the tremendous body of evidence showing the complex interplay among genes and environment in the pathophysiology of type 2 diabetes. Our group demonstrated how examination of such interactions could advance our knowledge of the underlying physiology of disease [51, 74, 83–86]. Indeed, the interaction between genes and environment may be one of the factors that will allow us to develop more personalized approaches to disease prevention. We already have evidence that genetic variation can affect drug response [87–91], and assessing interaction with environmental factors could help to narrow the focus to a smaller subset of individuals that might benefit from specifically tailored therapy.

The genetics of type 2 diabetes has seen a golden age of discovery that not only reinforced our knowledge of known biologic pathways, but also highlighted new biology underlying the disease. We are now entering a new age where this genetic knowledge must be translated to physiology and leveraged at the clinical level to improve management of the disease and reduce diabetes incidence worldwide. Genetics-based prevention could be the greatest contribution to disease research in decades. However, to achieve that goal, we need to better understand how genetic variation alters the underlying physiology that thrusts an individual onto the trajectory toward disease.

References

1. von Engelhardt D (1989) Diabetes: its medical and cultural history. Springer-Verlag, Berlin, Germany
2. Whiting DR, Guariguata L, Weil C, Shaw J (2011) IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes Res Clin Pract* 94:311–321
3. Shaw JE, Sicree RA, Zimmet PZ (2010) Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res Clin Pract* 87:4–14
4. The Diabetes Prevention Program Research G (2005) Prevention of type 2 diabetes with troglitazone in the diabetes prevention program. *Diabetes* 54:1150–1156
5. The DTI (2006) Effect of rosiglitazone on the frequency of diabetes in patients with impaired glucose tolerance or impaired fasting glucose: a randomised controlled trial. *Lancet* 368:1096–1105
6. Buchanan TA, Xiang AH, Peters RK, Kjos SL, Marroquin A, Goico J, Ochoa C, Tan S, Berkowitz K, Hodis HN, Azen SP (2002) Preservation of pancreatic β -cell function and prevention of type 2 diabetes by pharmacological treatment of insulin resistance in high-risk hispanic women. *Diabetes* 51:2796–2803
7. Xiang AH, Peters RK, Kjos SL, Marroquin A, Goico J, Ochoa C, Kawakubo M, Buchanan TA (2006) Effect of pioglitazone on pancreatic β -cell function and diabetes risk in Hispanic women with prior gestational diabetes. *Diabetes* 55:517–522
8. DeFronzo RA, Tripathy D, Schwenke DC, Banerji M, Bray GA, Buchanan TA, Clement SC, Henry RR, Hodis HN, Kitabchi AE, Mack WJ, Mudaliar S, Ratner RE, Williams K, Stentz FB, Musi N, Reaven PD (2011) Pioglitazone for diabetes prevention in impaired glucose tolerance. *N Engl J Med* 364:1104–1115
9. Neel JV (1962) Diabetes mellitus: a “thrifty” genotype rendered detrimental by “progress”? *Am J Hum Genet* 14:353–362
10. Neel JV (1976) Diabetes mellitus - a Geneticist's nightmare. In: Creutzfeldt W, Kobberling J, Neel JV (eds) *The genetics of diabetes*. Springer, New York, pp 1–11
11. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273:1516–1517
12. International Human Genome Mapping C (2001) A physical map of the human genome. *Nature* 409:934–941
13. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XB, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferreira S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejarival A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooshep S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Foslter C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S,

- Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. *Science* 291:1304–1351
14. The International HapMap Consortium (2003) The international HapMap project. *Nature* 426:789–796
 15. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, Novartis Institutes for Biomedical Research, Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson Boström K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Råstam L, Seliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjögren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumensiel B, Parkin M, DeFelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316:1331–1336
 16. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneally PM, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316:1341–1345
 17. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, JRB P, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney ASF, The Wellcome Trust Case Control C, McCarthy MI, Hattersley AT (2007) Replication of genome-wide association signals in U.K. samples reveals risk loci for type 2 diabetes. *Science* 316:1336–1341
 18. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P (2007) A genome-wide association study identified novel risk loci for type 2 diabetes. *Nature* 445:881–885
 19. Florez JC, Manning AK, Dupuis J, McAteer J, Irenze K, Gianniny L, Mirel DB, Fox CS, Cupples LA, Meigs JB (2007) A 100K genome-wide association scan for diabetes and related traits in the Framingham heart study. *Diabetes* 56:3063–3074
 20. Hayes MG, Pluzhnikov A, Miyake K, Sun Y, Ng MCY, Roe CA, Below JE, Nicolae RI, Konkashbaev A, Bell GI, Cox NJ, Hais CL (2007) Identification of type 2 diabetes genes in Mexican Americans through genome-wide association studies. *Diabetes* 56:3033–3044
 21. Hanson RL, Bogardus C, Duggan D, Kobes S, Knowlton M, Infante AM, Marovich L, Benitez D, Baier LJ, Knowler WC (2007) A search for variants associated with young-onset type 2 diabetes in American Indians in a 100K genotyping array. *Diabetes* 56:3045–3052
 22. Zeggini E, Scott LJ, Saxena R, Voight BF, Diabetes Genetics R, Meta-analysis C (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 40:638–645
 23. Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, Furuta H, Hirota Y, Mori H, Jonsson A, Sato Y, Yamagata K, Hinokio Y, Wang HY, Tanahashi T, Nakamura N, Oka Y, Iwasaki N, Iwamoto Y, Yamada Y, Seino Y, Maegawa H, Kashiwagi A, Takeda J, Maeda E, Shin HD, Cho YM, Park KS, Lee HK, Ng MC, Ma RC, So WY, Chan JC, Lyssenko V, Tuomi T, Nilsson P, Groop L, Kamatani N, Sekine A, Nakamura Y, Yamamoto K, Yoshida T, Tokunaga K, Itakura M, Makino H, Nanjo K, Kadowaki T, Kasuga M (2008) Variants in *KCNQ1* are associated with susceptibility to type 2 diabetes mellitus. *Nat Genet* 40:1092–1097
 24. Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, Andersen G, Ng DP, Holmkvist J, Borch-Johnsen K, Jorgensen T,

- Sandbaek A, Lauritzen T, Hansen T, Nurbaya S, Tsunoda T, Kubo M, Babazono T, Hirose H, Hayashi M, Iwamoto Y, Kashiwagi A, Kaku K, Kawamori R, Tai ES, Pedersen O, Kamatani N, Kadowaki T, Kikkawa R, Nakamura Y, Maeda S (2008) SNPs in *KCNQ1* are associated with susceptibility to type 2 diabetes in east Asian and European populations. *Nat Genet* 40:1098–1102
25. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segrè AV, van Hoek M, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, Boerwinkle E, Bonnycastle LL, Boström KB, Bravenboer B, Bumpstead S, Burtt NP, Charpentier G, Chines PS, Cornelis M, Couper DJ, Crawford G, Doney ASF, Elliott KS, Elliott AL, Erdos MR, Fox CS, Franklin CS, Ganser M, Gieger C, Grarup N, Green T, Griffin S, Groves CJ, Guiducci C, Hadjadj S, Hassanali N, Herder C, Isomaa B, Jackson AU, Johnson PRV, Jorgensen T, Kao WHL, Klopp N, Kong A, Kraft P, Kuusisto J, Lauritzen T, Li M, Lieveise A, Lindgren CM, Lyssenko V, Marre M, Meitinger T, Midthjell K, Morken MA, Narisu N, Nilsson P, Owen KR, Payne F, Perry JRB, Petersen AK, Platou C, Proença C, Prokopenko I, Rathmann W, Rayner NW, Robertson NR, Rocheleau G, Roden M, Sampson MJ, Saxena R, Shields BM, Shrader P, Sigurdsson G, Sparso T, Strassburger K, Stringham HM, Sun Q, Swift AJ, Thorand B, Tichet J, Tuomi T, van Dam RM, van Haeften TW, van Herpt T, van Vliet-Ostaptchouk JV, Walters GB, Weedon MN, Wijmenga C, Wittman J, The MI, The GC, Bergman RN, Cauchi S, Collins FS, Gloyn AL, Gyllensten U, Hansen T, Hide WA, Hitman GA, Hofman A, Hunter DJ, Hveem K, Laakso M, Mohlke KL, Morris AD, Palmer CNA, Pramstaller PP, Rudan I, Sijbrands E, Stein LD, Tuomilehto J, Uitterlinden A, Walker M, Wareham NJ, Watanabe RM, Abecasis GR, Boehm BO, Campbell H, Daly MJ, Hattersley AT, Hu FB, Meigs JB, Pankow JS, Pedersen O, Wichmann HE, Barroso I, Florez JC, Frayling TM, Groop L, Sladek R, Thorsteinsdottir U, Wilson JF, Illig T, Froguel P, Van Duijn CM, Stefansson K, Altshuler D, Boehnke M, McCarthy MI (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 42:579–589
26. Mahajan A, Go MJ, Zhang W, Below JE, Gaulton KJ, Ferreira T, Horikoshi M, Johnson AD, Ng MC, Prokopenko I, Saleheen D, Wang X, Zeggini E, Abecasis GR, Adair LS, Almgren P, Atalay M, Aung T, Baldassarre D, Balkau B, Bao Y, Barnett AH, Barroso I, Basit A, Been LF, Beilby J, Bell GI, Benediktsson R, Bergman RN, Boehm BO, Boerwinkle E, Bonnycastle LL, Burtt N, Cai Q, Campbell H, Carey J, Cauchi S, Caulfield M, Chan JC, Chang LC, Chang TJ, Chang YC, Charpentier G, Chen CH, Chen H, Chen YT, Chia KS, Chidambaram M, Chines PS, Cho NH, Cho YM, Chuang LM, Collins FS, Cornelis MC, Couper DJ, Crenshaw AT, van Dam RM, Danesh J, Das D, de Faire U, Dedoussis G, Deloukas P, Dimas AS, Dina C, Doney AS, Donnelly PJ, Dorkhan M, van Duijn C, Dupuis J, Edkins S, Elliott P, Emilsson V, Erbel R, Eriksson JG, Escobedo J, Esko T, Eury E, Florez JC, Fontanillas P, Forouhi NG, Forsen T, Fox C, Fraser RM, Frayling TM, Froguel P, Frossard P, Gao Y, Gertow K, Gieger C, Gigante B, Grallert H, Grant GB, Grrop LC, Groves CJ, Grundberg E, Guiducci C, Hamsten A, Han BG, Hara K, Hassanali N, Hattersley AT, Hayward C, Hedman AK, Herder C, Hofman A, Holmen OL, Hovingh K, Hreidarsson AB, Hu C, Hu FB, Hui J, Humphries SE, Hunt SE, Hunter DJ, Hveem K, Hydrie ZI, Ikegami H, Illig T, Ingelsson E, Islam M, Isomaa B, Jackson AU, Jafar T, James A, Jia W, Jockel KH, Jonsson A, Jowett JB, Kadowaki T, Kang HM, Kanoni S, Kao WH, Kathiresan S, Kato N, Katulanda P, Keinanen-Kiukaanniemi KM, Kelly AM, Khan H, Khaw KT, Khor CC, Kim HL, Kim S, Kim YJ, Kinnunen L, Klopp N, Kong A, Korpi-Hyovalti E, Kowlessur S, Kraft P, Kravic J, Kristensen MM, Krithika S, Kumar A, Kumate J, Kuusisto J, Kwak SH, Laakso M, Lagou V, Lakka TA, Langenberg C, Langford C, Lawrence R, Leander K, Lee JM, Lee NR, Li M, Li X, Li Y, Liang J, Liju S, Lim WY, Lind L, Lindgren CM, Lindholm E, Liu CT, Liu JJ, Lobbens S, Long J, Loos RJ, Lu W, Luan J, Lyssenko V, Ma RC, Maeda S, Magi R, Mannisto S, Matthews DR, Meigs JB, Melander O, Metspalu A, Meyer J, Mirza G, Mihailov E, Moebus S, Mohan V, Mohlke KL, Morris AD, Muhleisen TW, Muller-Nurasyid M, Musk B, Nakamura J, Nakashima E, Navarro P, Ng PK, Nica AC, Nilsson PM, Njolstad I, Nothen MM, Ohnaka K, Ong TH, Owen KR, Palmer CN, Pankow JS, Park KS, Parkin M, Pechlivanis S, Pedersen NL, Peltonen L, Perry JR, Peters A, Piniidiyapathirage JM, Platou CG, Potter S, Price JF, Qi L,

- Radha V, Rallidis L, Rasheed A, Rathman W, Rauramaa R, Raychaudhuri S, Rayner NW, Rees SD, Rehnberg E, Ripatti S, Robertson N, Roden M, Rossin EJ, Rudan I, Rybin D, Saaristo TE, Salomaa V, Saltevo J, Samuel M, Sanghera DK, Saramies J, Scott J, Scott LJ, Scott RA, Segre AV, Sehmi J, Sennblad B, Shah N, Shah S, Shera AS, Shu XO, Shuldiner AR, Sigurdsson G, Sijbrands E, Silveira A, Sim X, Sivapalaratnam S, Small KS, So WY, Stancakova A, Stefansson K, Steinbach G, Steinthorsdottir V, Stirrups K, Strawbridge RJ, Stringham HM, Sun Q, Suo C, Syvanen AC, Takayanagi R, Takeuchi F, Tay WT, Teslovich TM, Thorand B, Thorleifsson G, Thorsteinsdottir U, Tikkanen E, Trakalo J, Tremoli E, Trip MD, Tsai FJ, Tuomi T, Tuomilehto J, Uitterlinden AG, Valladares-Salgado A, Vedantam S, Veglia F, Voight BF, Wang C, Wareham NJ, Wennauer R, Wickremasinghe AR, Wilsgaard T, Wilson JF, Wiltshire S, Winckler W, Wong TY, Wood AR, Wu JY, Wu Y, Yamamoto K, Yamauchi T, Yang M, Yengo L, Yokota M, Young R, Zabaneh D, Zhang F, Zhang R, Zheng W, Zimmet PZ, Altshuler D, Bowden DW, Cho YS, Cox NJ, Cruz M, Hanis CL, Kooner J, Lee JY, Seielstad M, Teo YY, Boehnke M, Parra EJ, Chambers JC, Tai ES, McCarthy MI, Morris AP (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* 46:234–244
27. Scuteri A, Sanna S, Chen WM, Uda M, Albai G, Strait J, Najjar S, Nagaraja R, Orr \pounds M, Usala G, Dei M, Lai S, Maschio A, Busonero F, Mulas A, Ehret GB, Fink AA, Weder AB, Cooper RS, Galan P, Chakravarti A, Schlessinger D, Cao A, Lakatta E, Abecasis GR (2007) Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet* 3:1200–1210
28. Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S, Baker A, Snorrardottir S, Bjarnason H, Ng MCY, Hansen T, Bagger Y, Wilensky RL, Reilly MP, Adeyemo A, Chen Y, Zhou J, Gudnason V, Chen G, Huang H, Lashley K, Doumatey A, So WY, Ma RCY, Andersen G, Borch-Johnsen K, Jorgensen T, van Vliet-Ostaptchouk JV, Hofker MH, Wijmenga C, Christiansen C, Rader DJ, Rotimi C, Gurney M, Chan JCN, Pedersen O, Sigurdsson G, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet* 39:770–775
29. Chambers JC, Elliott P, Zabaneh D, Zhang W, Li Y, Froguel P, Balding D, Scott J, Kooner JS (2008) Common genetic variation near MC4R is associated with waist circumference and insulin resistance. *Nat Genet* 40:716–718
30. Chen WM, Erdos MR, Jackson AU, Saxena R, Sanna S, Silver KD, Timpson NJ, Hansen T, Orrù M, Piras MG, Bonnycastle LL, Willer CJ, Lyssenko V, Shen HL, Kuusisto J, Ebrahim S, Sestu N, Duren WL, Spada MC, Stringham HM, Scott LJ, Olla N, Swift AJ, Najjar S, Mitchell BD, Lawlor DA, Davey-Smith G, Ben-Shlomo Y, Andersen G, Borch-Johnsen K, Jorgensen T, Saramies J, Valle TT, Buchanan TA, Shuldiner AR, Lakatta E, Bergman RN, Uda M, Tuomilehto J, Pedersen O, Cao A, Groop L, Mohlke KL, Laakso M, Schlessinger D, Collins FS, Altshuler D, Abecasis GR, Boehnke M, Scuteri A, Watanabe RM (2008) Variations in the G6PC2/ABCB11 genomic region are associated with fasting glucose levels. *J Clin Invest* 118:2609–2628
31. Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, Inouye M, Freathy RM, Attwood AP, Beckmann JS, Berndt SI, Jacobs KB, Chanock SJ, Hayes RB, Bergmann S, Bennett AJ, Bingham SA, Bochud M, Brown M, Cauchi S, Connell JM, Cooper C, Smith GD, Day I, Dina C, De S, Dermizakis ET, Doney AS, Elliott KS, Elliott P, Evans DM, Sadaf FI, Froguel P, Ghorri J, Groves CJ, Gwilliam R, Hadley D, Hall AS, Hattersley AT, Hebebrand J, Heid IM, Lamina C, Gieger C, Illig T, Meitinger T, Wichmann HE, Herrera B, Hinney A, Hunt SE, Jarvelin MR, Johnson T, Jolley JD, Karpe F, Keniry A, Khaw KT, Luben RN, Mangino M, Marchini J, McArdle WL, McGinnis R, Meyre D, Munroe PB, Morris AD, Ness AR, Neville MJ, Nica AC, Ong KK, O'Rahilly S, Owen KR, Palmer CN, Papadakis K, Potter S, Pouta A, Qi L, Randall JC, Rayner NW, Ring SM, Sandhu MS, Scherag A, Sims MA, Song K, Soranzo N, Speliotes EK, Syddall HE, Teichmann SA, Timpson NJ, Tobias JH, Uda M, Vogel CI, Wallace C, Waterworth DM, Weedon MN, Willer CJ, Wraight YX, Zeggini E, Hirschhorn JN, Strachan DP, Ouwehand WH, Caulfield MJ, Samani NJ, Frayling TM, Vollenweider P, Waeber G, Mooser V, Deloukas P, McCarthy MI, Wareham NJ, Barroso I, Jacobs KB, Chanock SJ, Hayes RB, Lamina C, Gieger C, Illig T, Meitinger T, Wichmann HE, Kraft P, Hankinson SE, Hunter DJ, Hu FB, Lyon HN, Voight BF, Ridderstrale M, Groop L, Scheet P, Sanna S, Abecasis GR, Albai G, Nagaraja R, Schlessinger D, Jackson AU, Tuomilehto J,

- Collins FS, Boehnke M, Mohlke KL (2008) Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet* 40:768–775
32. Orho-Melander M, Melander O, Guiducci C, Perez-Martinez P, Corella D, Roos C, Tewhey R, Rieder MJ, Hall J, Abecasis G, Tai ES, Welch C, Arnett DK, Lyssenko V, Lindholm E, Saxena R, de Bakker PI, Burt N, Voight BF, Hirschhorn JN, Tucker KL, Hedner T, Tuomi T, Isomaa B, Eriksson KF, Taskinen MR, Wahlstrand B, Hughes TE, Parnell LD, Lai CQ, Berglund G, Peltonen L, Vartiainen E, Jousilahti P, Havulinna AS, Salomaa V, Nilsson P, Groop L, Altshuler D, Ordovas JM, Kathiresan S (2008) Common missense variant in the glucokinase regulatory protein gene is associated with increased plasma triglyceride and C-reactive protein but lower fasting glucose concentrations. *Diabetes* 57:3112–3121
33. Pare G, Chasman DI, Parker AN, Nathan DM, Miletich JP, Zee RY, Ridker PM (2008) Novel association of HK1 with glycated hemoglobin in a non-diabetic population: a genome-wide evaluation of 14,618 participants in the Women's genome health study. *PLoS Genet* 4:e1000312
34. Romeo S, Kozlitina J, Xing C, Pertsemlidis A, Cox D, Pennacchio LA, Boerwinkle E, Cohen JC, Hobbs HH (2008) Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* 40:1461–1465
35. Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, Bonnycastle LL, Shen H, Timpson N, Lettre G, Usala G, Chines PS, Stringham HM, Scott LJ, Dei M, Lai S, Albai G, Crisponi L, Naitza S, Doheny KF, Pugh EW, Ben-Shlomo Y, Ebrahim S, Lawlor DA, Bergman RN, Watanabe RM, Uda M, Tuomilehto J, Coresh J, Hirschhorn JN, Shuldiner AR, Schlessinger D, Collins FS, Davey Smith G, Boerwinkle E, Cao A, Boehnke M, Abecasis GR, Mohlke KL (2008) Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat Genet* 40:198–203
36. Beer NL, Tribble ND, McCulloch LJ, Roos C, Johnson PR, Orho-Melander M, Gloyn AL (2009) The P446L variant in GCKR associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. *Hum Mol Genet* 18:4081–4088
37. Bouatia-Naji N, Bonnefond A, Cavalcanti-Proenca C, Sparso T, Holmkvist J, Marchand M, Delplanque J, Lobbens S, Rocheleau G, Durand E, De GF, Chevre JC, Borch-Johnsen K, Hartikainen AL, Ruokonen A, Tichet J, Marre M, Weill J, Heude B, Tauber M, Lemaire K, Schuit F, Elliott P, Jorgensen T, Charpentier G, Hadjadj S, Cauchi S, Vaxillaire M, Sladek R, Visvikis-Siest S, Balkau B, Levy-Marchal C, Pattou F, Meyre D, Blakemore AI, Jarvelin MR, Walley AJ, Hansen T, Dina C, Pedersen O, Froguel P (2009) A variant near MTNR1B is associated with increased fasting plasma glucose levels and type 2 diabetes risk. *Nat Genet* 41:89–94
38. Lindgren CM, Heid IM, Randall JC, Lamina C, Steinthorsdottir V, Qi L, Speliotes EK, Thorleifsson G, Willer CJ, Herrera BM, Jackson AU, Lim N, Scheet P, Soranzo N, Amin N, Aulchenko YS, Chambers JC, Drong A, Luan J, Lyon HN, Rivadeneira F, Sanna S, Timpson NJ, Zillikens MC, Zhao JH, Almgren P, Bandinelli S, Bennett AJ, Bergman RN, Bonnycastle LL, Bumpstead SJ, Chanoock SJ, Cherkas L, Chines P, Coin L, Cooper C, Crawford G, Doering A, Dominiczak A, Doney AS, Ebrahim S, Elliott P, Erdos MR, Estrada K, Ferrucci L, Fischer G, Forouhi NG, Gieger C, Grallert H, Groves CJ, Grundy S, Guiducci C, Hadley D, Hamsten A, Havulinna AS, Hofman A, Holle R, Holloway JW, Illig T, Isomaa B, Jacobs LC, Jameson K, Jousilahti P, Karpe F, Kuusisto J, Laitinen J, Lathrop GM, Lawlor DA, Mangino M, McArdle WL, Meitinger T, Morken MA, Morris AP, Munroe P, Narisu N, Nordstrom A, Nordstrom P, Oostra BA, Palmer CN, Payne F, Peden JF, Prokopenko I, Renstrom F, Ruokonen A, Salomaa V, Sandhu MS, Scott LJ, Scuteri A, Silander K, Song K, Yuan X, Stringham HM, Swift AJ, Tuomi T, Uda M, Vollenweider P, Waeber G, Wallace C, Walters GB, Weedon MN, Wittteman JC, Zhang C, Zhang W, Caulfield MJ, Collins FS, Davey SG, Day IN, Franks PW, Hattersley AT, FB H, Jarvelin MR, Kong A, Kooner JS, Laakso M, Lakatta E, Mooser V, Morris AD, Peltonen L, Samani NJ, Spector TD, Strachan DP, Tanaka T, Tuomilehto J, Uitterlinden AG, Van Duijn CM, Wareham NJ, Hugh W, Waterworth DM, Boehnke M, Deloukas P, Groop L, Hunter DJ, Thorsteinsdottir U, Schlessinger D, Wichmann HE, Frayling TM, Abecasis GR, Hirschhorn JN, Loos RJ, Stefansson K, Mohlke KL, Barroso I, McCarthy MI (2009) Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genet* 5:e1000508
39. Lyssenko V, Nagorny CL, Erdos MR, Wierup N, Jonsson A, Spiegel P, Bugliani M,

- Saxena R, Fex M, Pulizzi N, Isomaa B, Tuomi T, Nilsson P, Kuusisto J, Tuomilehto J, Boehnke M, Altshuler D, Sundler F, Eriksson JG, Jackson AU, Laakso M, Marchetti P, Watanabe RM, Mulder H, Groop L (2009) Common variant in *MTNR1B* associated with increased risk of type 2 diabetes and impaired early insulin secretion. *Nat Genet* 41:82–88
40. Prokopenko I, Langenberg C, Florez JC, Saxena R, Soranzo N, Thorleifsson G, Loos RJ, Manning AK, Jackson AU, Aulchenko Y, Potter SC, Erdos MR, Sanna S, Hottenga JJ, Wheeler E, Kaakinen M, Lyssenko V, Chen WM, Ahmadi K, Beckmann JS, Bergman RN, Bochud M, Bonnycastle LL, Buchanan TA, Cao A, Cervino A, Coin L, Collins FS, Crisponi L, de Geus EJ, Dehghan A, Deloukas P, Doney AS, Elliott P, Freimer N, Gateva V, Herder C, Hofman A, Hughes TE, Hunt S, Illig T, Inouye M, Isomaa B, Johnson T, Kong A, Krestyaninova M, Kuusisto J, Laakso M, Lim N, Lindblad U, Lindgren CM, McCann OT, Mohlke KL, Morris AD, Naitza S, Orru M, Palmer CN, Pouta A, Randall J, Rathmann W, Saramies J, Scheet P, Scott LJ, Scuteri A, Sharp S, Sijbrands E, Smit JH, Song K, Steinthorsdottir V, Stringham HM, Tuomi T, Tuomilehto J, Uitterlinden AG, Voight BF, Waterworth D, Wichmann HE, Willemsen G, Witteman JC, Yuan X, Zhao JH, Zeggini E, Schlessinger D, Sandhu M, Boomsma DI, Uda M, Spector TD, Penninx BW, Altshuler D, Vollenweider P, Jarvelin MR, Lakatta E, Waeber G, Fox CS, Peltonen L, Groop LC, Mooser V, Cupples LA, Thorsteinsdottir U, Boehnke M, Barroso I, Van DC, Dupuis J, Watanabe RM, Stefansson K, McCarthy MI, Wareham NJ, Meigs JB, Abecasis GR (2009) Variants in *MTNR1B* influence fasting glucose levels. *Nat Genet* 41:77–81
41. Rung J, Cauchi S, Albrechtsen A, Shen L, Rocheleau G, Cavalcanti-Proenca C, Bacot F, Balkau B, Belisle A, Borch-Johnsen K, Charpentier G, Dina C, Durand E, Elliott P, Hadjadj S, Jarvelin MR, Laitinen J, Lauritzen T, Marre M, Mazur A, Meyre D, Montpetit A, Pisinger C, Posner B, Poulsen P, Pouta A, Prentki M, Ribel-Madsen R, Ruokonen A, Sandbaek A, Serre D, Tichet J, Vaxillaire M, Wojtaszewski JF, Vaag A, Hansen T, Polychronakos C, Pedersen O, Froguel P, Sladek R (2009) Genetic variant near *IRS1* is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nat Genet* 41:1110–1115
42. Willer CJ, Speliotes EK, Loos RJ, Li S, Lindgren CM, Heid IM, Berndt SI, Elliott AL, Jackson AU, Lamina C, Lettre G, Lim N, Lyon HN, McCarroll SA, Papadakis K, Qi L, Randall JC, Roccacella RM, Sanna S, Scheet P, Weedon MN, Wheeler E, Zhao JH, Jacobs LC, Prokopenko I, Soranzo N, Tanaka T, Timpson NJ, Almgren P, Bennett A, Bergman RN, Bingham SA, Bonnycastle LL, Brown M, Burtt NP, Chines P, Coin L, Collins FS, Connell JM, Cooper C, Smith GD, Dennison EM, Deodhar P, Elliott P, Erdos MR, Estrada K, Evans DM, Gianniny L, Gieger C, Gillson CJ, Guiducci C, Hackett R, Hadley D, Hall AS, Havulinna AS, Hebebrand J, Hofman A, Isomaa B, Jacobs KB, Johnson T, Jousilahti P, Jovanovic Z, Khaw KT, Kraft P, Kuokkanen M, Kuusisto J, Laitinen J, Lakatta EG, Luan J, Luben RN, Mangino M, McArdle WL, Meitinger T, Mulas A, Munroe PB, Narisu N, Ness AR, Northstone K, O’Rahilly S, Purmann C, Rees MG, Ridderstrale M, Ring SM, Rivadeneira F, Ruokonen A, Sandhu MS, Saramies J, Scott LJ, Scuteri A, Silander K, Sims MA, Song K, Stephens J, Stevens S, Stringham HM, Tung YC, Valle TT, Van Duijn CM, Vimalaswaran KS, Vollenweider P, Waeber G, Wallace C, Watanabe RM, Waterworth DM, Watkins N, Witteman JC, Zeggini E, Zhai G, Zillikens MC, Altshuler D, Caulfield MJ, Chanoock SJ, Farooqi IS, Ferrucci L, Guralnik JM, Hattersley AT, FB H, Jarvelin MR, Laakso M, Mooser V, Ong KK, Ouwehand WH, Salomaa V, Samani NJ, Spector TD, Tuomi T, Tuomilehto J, Uda M, Uitterlinden AG, Wareham NJ, Deloukas P, Frayling TM, Groop LC, Hayes RB, Hunter DJ, Mohlke KL, Peltonen L, Schlessinger D, Strachan DP, Wichmann HE, McCarthy MI, Boehnke M, Barroso I, Abecasis GR, Hirschhorn JN (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* 41:25–34
43. Zhao J, Li M, Bradfield JP, Wang K, Zhang H, Sleiman P, Kim CE, Annaiah K, Glaberson W, Glessner JT, Otieno FG, Thomas KA, Garris M, Hou C, Frackelton EC, Chiavacci RM, Berkowitz RI, Hakonarson H, Grant SF (2009) Examination of type 2 diabetes loci implicates *CDKAL1* as a birth weight gene. *Diabetes* 58:2414–2418
44. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, Wheeler E, Glazer NL, Bouatia-Naji N, Gloy AL, Lindgren CM, Magi R, Morris AP, Randall J, Johnson T, Elliott P, Rybin D, Thorleifsson G, Steinthorsdottir V, Henneman P, Grallert H, Dehghan A, Hottenga JJ, Franklin CS, Navarro P, Song K,

- Goel A, Perry JR, Egan JM, Lajunen T, Grarup N, Sparso T, Doney A, Voight BF, Stringham HM, Li M, Kanoni S, Shrader P, Cavalcanti-Proenca C, Kumari M, Qi L, Timpson NJ, Gieger C, Zabena C, Rocheleau G, Ingelsson E, An P, O'Connell J, Luan J, Elliott A, McCarroll SA, Payne F, Roccascocca RM, Pattou F, Sethupathy P, Ardlie K, Ariyurek Y, Balkau B, Barter P, Beilby JP, Ben-Shlomo Y, Benediktsson R, Bennett AJ, Bergmann S, Bochud M, Boerwinkle E, Bonnefond A, Bonnycastle LL, Borch-Johnsen K, Bottcher Y, Brunner E, Bumpstead SJ, Charpentier G, Chen YD, Chines P, Clarke R, Coin LJ, Cooper MN, Cornelis M, Crawford G, Crisponi L, Day IN, de Geus EJ, Delplanque J, Dina C, Erdos MR, Fedson AC, Fischer-Rosinsky A, Forouhi NG, Fox CS, Frants R, Franzosi MG, Galan P, Goodarzi MO, Graessler J, Groves CJ, Grundy S, Gwilliam R, Gyllenstein U, Hadjadj S, Hallmans G, Hammond N, Han X, Hartikainen AL, Hassanali N, Hayward C, Heath SC, Hercberg S, Herder C, Hicks AA, Hillman DR, Hingorani AD, Hofman A, Hui J, Hung J, Isomaa B, Johnson PR, Jorgensen T, Julia A, Kaakinen M, Kaprio J, Kesaniemi YA, Kivimaki M, Knight B, Koskinen S, Kovacs P, Kyvik KO, Lathrop GM, Lawlor DA, Le BO, Lecoeur C, Li Y, Lyssenko V, Mahley R, Mangino M, Manning AK, Martinez-Larrad MT, McAteer JB, McCulloch LJ, McPherson R, Meisinger C, Melzer D, Meyre D, Mitchell BD, Morken MA, Mukherjee S, Naitza S, Narisu N, Neville MJ, Oostra BA, Orru M, Pakyz R, Palmer CN, Paoiliso G, Pattaro C, Pearson D, Peden JF, Pedersen NL, Perola M, Pfeiffer AF, Pichler I, Polasek O, Posthuma D, Potter SC, Pouta A, Province MA, Psaty BM, Rathmann W, Rayner NW, Rice K, Ripatti S, Rivadeneira F, Roden M, Rolandsson O, Sandbaek A, Sandhu M, Sanna S, Sayer AA, Scheet P, Scott LJ, Seedorf U, Sharp SJ, Shields B, Sigurethsson G, Sijbrands EJ, Silveira A, Simpson L, Singleton A, Smith NL, Sovio U, Swift A, Syddall H, Syvanen AC, Tanaka T, Thorand B, Tichet J, Tonjes A, Tuomi T, Uitterlinden AG, van Dijk KW, van HM, Varma D, Visvikis-Siest S, Vitart V, Vogelzangs N, Waeber G, Wagner PJ, Walley A, Walters GB, Ward KL, Watkins H, Weedon MN, Wild SH, Willemsen G, Wittmann JC, Yarnell JW, Zeggini E, Zelenika D, Zethelius B, Zhai G, Zhao JH, Zillikens MC, Borecki IB, Loos RJ, Meneton P, Magnusson PK, Nathan DM, Williams GH, Hattersley AT, Silander K, Salomaa V, Smith GD, Bornstein SR, Schwarz P, Spranger J, Karpe F, Shuldiner AR, Cooper C, Dedoussis GV, Serrano-Rios M, Morris AD, Lind L, Palmer LJ, Hu FB, Franks PW, Ebrahim S, Marmot M, Kao WH, Pankow JS, Sampson MJ, Kuusisto J, Laakso M, Hansen T, Pedersen O, Pramstaller PP, Wichmann HE, Illig T, Rudan I, Wright AF, Stumvoll M, Campbell H, Wilson JF, Hamsten APC, Investigators M, Bergman RN, Buchanan TA, Collins FS, Mohlke KL, Tuomilehto J, Valle TT, Altshuler D, Rotter JI, Siscovick DS, Penninx BW, Boomsma DI, Deloukas P, Spector TD, Frayling TM, Ferrucci L, Kong A, Thorsteinsdottir U, Stefansson K, van Duijn CM, Aulchenko YS, Cao A, Scuteri A, Schlessinger D, Uda M, Ruokonen A, Jarvelin MR, Waterworth DM, Vollenweider P, Peltonen L, Mooser V, Abecasis GR, Wareham NJ, Sladek R, Froguel P, Watanabe RM, Meigs JB, Groop L, Boehnke M, McCarthy MI, Florez JC, Barroso I (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* 42:105–116
45. Heid IM, Jackson AU, Randall JC, Winkler TW, Qi L, Steinthorsdottir V, Thorleifsson G, Zillikens MC, Speliotes EK, Magi R, Workalemahu T, White CC, Bouatia-Naji N, Harris TB, Berndt SI, Ingelsson E, Willer CJ, Weedon MN, Luan J, Vedantam S, Esko T, Kilpelainen TO, Kutalik Z, Li S, Monda KL, Dixon AL, Holmes CC, Kaplan LM, Liang L, Min JL, Moffatt MF, Molony C, Nicholson G, Schadt EE, Zondervan KT, Feitosa MF, Ferreira T, Allen HL, Weyant RJ, Wheeler E, Wood AR, Estrada K, Goddard ME, Lettre G, Mangino M, Nyholt DR, Purcell S, Smith AV, Visscher PM, Yang J, McCarroll SA, Nemesh J, Voight BF, Absher D, Amin N, Aspelund T, Coin L, Glazer NL, Hayward C, Heard-Costa NL, Hottenga JJ, Johansson A, Johnson T, Kaakinen M, Kapur K, Ketkar S, Knowles JW, Kraft P, Kraja AT, Lamina C, Leitzmann MF, McKnight B, Morris AP, Ong KK, Perry JR, Peters MJ, Polasek O, Prokopenko I, Rayner NW, Ripatti S, Rivadeneira F, Robertson NR, Sanna S, Sovio U, Surakka I, Teumer A, van WS, Vitart V, Zhao JH, Cavalcanti-Proenca C, Chines PS, Fisher E, Kulzer JR, Lecoeur C, Narisu N, Sandholt C, Scott LJ, Silander K, Stark K, Tammesoo ML, Teslovich TM, Timpson NJ, Watanabe RM, Welch R, Chasman DI, Cooper MN, Jansson JO, Kettunen J, Lawrence RW, Pellikka N, Perola M, Vandenput L, Alavere H, Almgren P, Atwood LD, Bennett AJ, Biffar R, Bonnycastle LL, Bornstein SR, Buchanan TA, Campbell H, Day IN, Dei M, Dorr M, Elliott P, Erdos MR, Eriksson JG, Freimer NB, Fu M, Gaget S, Geus EJ, Gjesing AP, Grallert H, Grassler J, Groves CJ,

- Guiducci C, Hartikainen AL, Hassanali N, Havulinna AS, Herzig KH, Hicks AA, Hui J, Igl W, Jousilahti P, Jula A, Kajantie E, Kinnunen L, Kolcic I, Koskinen S, Kovacs P, Kroemer HK, Krzelj V, Kuusisto J, Kvaloy K, Laitinen J, Lantieri O, Lathrop GM, Lokki ML, Luben RN, Ludwig B, McArdle WL, McCarthy A, Morken MA, Nelis M, Neville MJ, Pare G, Parker AN, Peden JF, Pichler I, Pietilainen KH, Platou CG, Pouta A, Ridderstrale M, Samani NJ, Saramies J, Sinisalo J, Smit JH, Strawbridge RJ, Stringham HM, Swift AJ, Teder-Laving M, Thomson B, Usala G, van Meurs JB, van Ommen GJ, Vatin V, Volpato CB, Wallaschofski H, Walters GB, Widen E, Wild SH, Willemssen G, Witte DR, Zgaga L, Zitting P, Beilby JP, James AL, Kahonen M, Lehtimäki T, Nieminen MS, Ohlsson C, Palmer LJ, Raitakari O, Ridker PM, Stumvoll M, Tonjes A, Viikari J, Balkau B, Ben-Shlomo Y, Bergman RN, Boeing H, Smith GD, Ebrahim S, Froguel P, Hansen T, Hengstenberg C, Hveem K, Isomaa B, Jorgensen T, Karpe F, Khaw KT, Laakso M, Lawlor DA, Marre M, Meitinger T, Metspalu A, Midthjell K, Pedersen O, Salomaa V, Schwarz PE, Tuomi T, Tuomilehto J, Valle TT, Wareham NJ, Arnold AM, Beckmann JS, Bergmann S, Boerwinkle E, Boomsma DI, Caulfield MJ, Collins FS, Eiriksdottir G, Gudnason V, Gyllenstein U, Hamsten A, Hattersley AT, Hofman A, Hu FB, Illig T, Iribarren C, Jarvelin MR, Kao WH, Kaprio J, Launer LJ, Munroe PB, Oostra B, Penninx BW, Pramstaller PP, Psaty BM, Quertermous T, Rissanen A, Rudan I, Shuldiner AR, Soranzo N, Spector TD, Syvanen AC, Uda M, Uitterlinden A, Volzke H, Vollenweider P, Wilson JF, Witteman JC, Wright AF, Abecasis GR, Boehnke M, Borecki IB, Deloukas P, Frayling TM, Groop LC, Haritunians T, Hunter DJ, Kaplan RC, North KE, O'Connell JR, Peltonen L, Schlessinger D, Strachan DP, hirschhorn JN, Assimes TL, Wichmann HE, Thorsteinsdottir U, Van Duijn CM, Stefansson K, Cupples LA, Loos RJ, Barroso I, McCarthy MI, Fox CS, Mohlke KL, Lindgren CM (2010) Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet* 42:960
46. Saxena R, Hivert MF, Langenberg C, Tanaka T, Pankow JS, Vollenweider P, Lyssenko V, Bouatia-Naji N, Dupuis J, Jackson AU, Kao WH, Li M, Glazer NL, Manning AK, Luan J, Stringham HM, Prokopenko I, Johnson T, Grarup N, Boesgaard TW, Lecoecur C, Shrader P, O'Connell J, Ingelsson E, Couper DJ, Rice K, Song K, Andreassen CH, Dina C, Kottgen A, Le BO, Pattou F, Taneera J, Steinthorsdottir V, Rybin D, Ardlie K, Sampson M, Qi L, van HM, Weedon MN, Aulchenko YS, Voight BF, Grallert H, Balkau B, Bergman RN, Bielinski SJ, Bonnefond A, Bonnycastle LL, Borch-Johnsen K, Bottcher Y, Brunner E, Buchanan TA, Bumpstead SJ, Cavalcanti-Proenca C, Charpentier G, Chen YD, Chines PS, Collins FS, Cornelis M, Crawford J, Delplanque J, Doney A, Egan JM, Erdos MR, Firmann M, Forouhi NG, Fox CS, Goodarzi MO, Graessler J, Hingorani A, Isomaa B, Jorgensen T, Kivimäki M, Kovacs P, Krohn K, Kumari M, Lauritzen T, Levy-Marchal C, May- or V, McAteer JB, Meyre D, Mitchell BD, Mohlke KL, Morken MA, Narisu N, Palmer CN, Pakyz R, Pascoe L, Payne F, Pearson D, Rathmann W, Sandbaek A, Sayer AA, Scott LJ, Sharp SJ, Sijbrands E, Singleton A, Siscovick DS, Smith NL, Sparso T, Swift AJ, Syddall H, Thorleifsson G, Tonjes A, Tuomi T, Tuomilehto J, Valle TT, Waeber G, Walley A, Waterworth DM, Zeggini E, Zhao JH, Illig T, Wichmann HE, Wilson JF, Van DC, Hu FB, Morris AD, Frayling TM, Hattersley AT, Thorsteinsdottir U, Stefansson K, Nilsson P, Syvanen AC, Shuldiner AR, Walker M, Bornstein SR, Schwarz P, Williams GH, Nathan DM, Kuusisto J, Laakso M, Cooper C, Marmot M, Ferrucci L, Mooser V, Stumvoll M, Loos RJ, Altshuler D, Psaty BM, Rotter JI, Boerwinkle E, Hansen T, Pedersen O, Florez JC, McCarthy MI, Boehnke M, Barroso I, Sladek R, Froguel P, Meigs JB, Groop L, Wareham NJ, Watanabe RM (2010) Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet* 42:142–148
47. Strawbridge RJ, Dupuis J, Prokopenko I, Barker A, Ahlqvist E, Rybin D, Petrie JR, Travers ME, Bouatia-Naji N, Dimas AS, Nica A, Wheeler E, Chen H, Voight BF, Taneera J, Kanoni S, Peden JF, Turrini F, Gustafsson S, Zabena C, Almgren P, Barker DJ, Barnes D, Dennison EM, Eriksson JG, Eriksson P, Eury E, Folkersen L, Fox CS, Frayling TM, Goel A, Gu HF, Horikoshi M, Isomaa B, Jackson AU, Jameson KA, Kajantie E, Kerr-Conte J, Kuulasmaa T, Kuusisto J, Loos RJ, Luan J, Makrilakis K, Manning AK, Martinez-Larrad MT, Narisu N, Nastase MM, Ohrvik J, Osmond C, Pascoe L, Payne F, Sayer AA, Sennblad B, Silveira A, Stancakova A, Stirrups K, Swift AJ, Syvanen AC, Tuomi T, van 't Hooft FM, Walker M, Weedon MN, Xie W, Zethelius B, Ongen H, Malarstig A, Hopewell JC, Saleheen D, Chambers J,

- Parish S, Danesh J, Kooner J, Ostenson CG, Lind L, Cooper CC, Serrano-Rios M, Ferrannini E, Forsen TJ, Clarke R, Franzosi MG, Seedorf U, Watkins H, Froguel P, Johnson P, Deloukas P, Collins FS, Laakso M, Dermitzakis ET, Boehnke M, McCarthy MI, Wareham NJ, Groop L, Pattou F, Gloyn AL, Dedoussis GV, Lyssenko V, Meigs JB, Barroso I, Watanabe RM, Ingelsson E, Langenberg C, Hamsten A, Florez JC (2011) Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* 60:2624–2634
48. Manning AK, Hivert MF, Scott RA, Grimsby JL, Bouatia-Naji N, Chen H, Rybin D, Liu CT, Bielak LF, Prokopenko I, Amin N, Barnes D, Cadby G, Hottenga JJ, Ingelsson E, Jackson AU, Johnson T, Kanoni S, Ladenvall C, Lagou V, Lahti J, Lecocour C, Liu Y, Martinez-Larrad MT, Montasser ME, Navarro P, Perry JR, Rasmussen-Torvik LJ, Salo P, Sattar N, Shungin D, Strawbridge RJ, Tanaka T, Van Duijn CM, An P, de AM, Andrews JS, Aspelund T, Atalay M, Aulchenko Y, Balkau B, Bandinelli S, Beckmann JS, Beilby JP, Bellis C, Bergman RN, Blangero J, Boban M, Boehnke M, Boerwinkle E, Bonnycastle LL, Boomsma DI, Borecki IB, Bottcher Y, Bouchard C, Brunner E, Budimir D, Campbell H, Carlson O, Chines PS, Clarke R, Collins FS, Corbaton-Anchuelo A, Couper D, de Faire U, Dedoussis GV, Deloukas P, Dimitriou M, Egan JM, Eiriksdottir G, Erdos MR, Eriksson JG, Eury E, Ferrucci L, Ford I, Forouhi NG, Fox CS, Franzosi MG, Franks PW, Frayling TM, Froguel P, Galan P, de GE, Gigante B, Glazer NL, Goel A, Groop L, Gudnason V, Hallmans G, Hamsten A, Hansson O, Harris TB, Hayward C, Heath S, Hercberg S, Hicks AA, Hingorani A, Hofman A, Hui J, Hung J, Jarvelin MR, Jhun MA, Johnson PC, Jukema JW, Jula A, Kao WH, Kaprio J, Kardina SL, Keinanen-Kiukkaanniemi S, Kivimaki M, Kolcic I, Kovacs P, Kumari M, Kuusisto J, Kyvik KO, Laakso M, Lakka T, Lannfelt L, Lathrop GM, Launer LJ, Leander K, Li G, Lind L, Lindstrom J, Lobbens S, Loos RJ, Luan J, Lyssenko V, Magi R, Magnusson PK, Marmot M, Meneton P, Mohlke KL, Mooser V, Morken MA, Miljkovic I, Narisu N, O'Connell J, Ong KK, Oostra BA, Palmer LJ, Palotie A, Pankow JS, Peden JF, Pedersen NL, Pehlic M, Peltonen L, Penninx B, Pericic M, Perola M, Perusse L, Peyser PA, Polasek O, Pramstaller PP, Province MA, Raikonen K, Rauramaa R, Rehnberg E, Rice K, Rotter JI, Rudan I, Ruukonen A, Saaristo T, Sabater-Lleal M, Salomaa V, Savage DB, Saxena R, Schwarz P, Seedorf U, Sennblad B, Serrano-Rios M, Shuldiner AR, Sijbrands EJ, Siscovick DS, Smit JH, Small KS, Smith NL, Smith AV, Stancakova A, Stirrups K, Stumvoll M, Sun YV, Swift AJ, Tonjes A, Tuomilehto J, Trompet S, Uitterlinden AG, Uusitupa M, Vikstrom M, Vitart V, Vohl MC, Voight BF, Vollenweider P, Waeber G, Waterworth DM, Watkins H, Wheeler E, Widen E, Wild SH, Willems SM, Willemsen G, Wilson JF, Witteman JC, Wright AF, Yaghootkar H, Zelenika D, Zemunik T, Zgaga L, Wareham NJ, McCarthy MI, Barroso I, Watanabe RM, Florez JC, Dupuis J, Meigs JB, Langenberg C (2012) A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* 44:659–669
49. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. *Nature* 461:747–753
50. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, Ma C, Fontanillas P, Moutsianas L, McCarthy DJ, Rivas MA, Perry JR, Sim X, Blackwell TW, Robertson NR, Rayner NW, Cingolani P, Locke AE, Fernandez Tajos J, Highland HM, Dupuis J, Chines PS, Lindgren CM, Hartl C, Jackson AU, Chen H, Huyghe JR, van de Bunt M, Pearson RD, Kumar A, Muller-Nurasyid M, Grarup N, Stringham HM, Gamazon ER, Lee J, Chen Y, Scott RA, Below JE, Chen P, Huang J, Go MJ, Stitzel ML, Pasko D, Parker SC, Varga TV, Green T, Beer NL, Day-Williams AG, Ferreira T, Fingerlin T, Horikoshi M, Hu C, Huh I, Ikram MK, Kim BJ, Kim Y, Kim YJ, Kwon MS, Lee J, Lee S, Lin KH, Maxwell TJ, Nagai Y, Wang X, Welch RP, Yoon J, Zhang W, Barzilay N, Voight BF, Han BG, Jenkinson CP, Kuulasmaa T, Kuusisto J, Manning A, Ng MC, Palmer ND, Balkau B, Stancakova A, Abboud HE, Boeing H, Giedraitis V, Prabhakaran D, Gottesman O, Scott J, Carey J, Kwan P, Grant G, Smith JD, Neale BM, Purcell S, Butterworth AS, Howson JM, Lee HM, Lu Y, Kwak SH, Zhao W, Danesh J, Lam VK, Park KS, Saleheen D, So WY, Tam CH, Afzal U, Aguilar D, Arya R,

- Aung T, Chan E, Navarro C, Cheng CY, Palli D, Correa A, Curran JE, Rybin D, Farook VS, Fowler SP, Freedman BI, Griswold M, Hale DE, Hicks PJ, Khor CC, Kumar S, Lehne B, Thuillier D, Lim WY, Liu J, van der Schouw YT, Loh M, Musani SK, Puppala S, Scott WR, Yengo L, Tan ST, Taylor HA Jr, Thameem F, Wilson G Sr, Wong TY, Njolstad PR, Levy JC, Mangino M, Bonnycastle LL, Schwarzmayr T, Fadista J, Surdulescu GL, Herder C, Groves CJ, Wieland T, Bork-Jensen J, Brandslund I, Christensen C, Koistinen HA, Doney AS, Kinnunen L, Esko T, Farmer AJ, Hakaste L, Hodgkiss D, Kravic J, Lyssenko V, Hollensted M, Jorgensen ME, Jorgensen T, Ladenvall C, Justesen JM, Karajamaki A, Kriebel J, Rathmann W, Lannfelt L, Lauritzen T, Narisu N, Linneberg A, Melander O, Milani L, Neville M, Orholm-Melander M, Qi L, Qi Q, Roden M, Rolandsson O, Swift A, Rosengren AH, Stirrups K, Wood AR, Mihailov E, Blancher C, Carneiro MO, Maguire J, Poplin R, Shakir K, Fennell T, DePristo M, Hrabe de Angelis M, Deloukas P, Gjesing AP, Jun G, Nilsson P, Murphy J, Onofrio R, Thorand B, Hansen T, Meisinger C, Hu FB, Isomaa B, Karpe F, Liang L, Peters A, Huth C, O'Rahilly SP, Palmer CN, Pedersen O, Rauramaa R, Tuomilehto J, Salomaa V, Watanabe RM, Syvanen AC, Bergman RN, Bharadwaj D, Bottinger EP, Cho YS, Chandak GR, Chan JC, Chia KS, Daly MJ, Ebrahim SB, Langenberg C, Elliott P, Jablonski KA, Lehman DM, Jia W, Ma RC, Pollin TI, Sandhu M, Tandon N, Froguel P, Barroso I, Teo YY, Zeggini E, Loos RJ, Small KS, Ried JS, DeFronzo RA, Grallert H, Glaser B, Metspalu A, Wareham NJ, Walker M, Banks E, Gieger C, Ingelsson E, Im HK, Illig T, Franks PW, Buck G, Trakalo J, Buck D, Prokopenko I, Magi R, Lind L, Farjoun Y, Owen KR, Gloyn AL, Strauch K, Tuomi T, Kooner JS, Lee JY, Park T, Donnelly P, Morris AD, Hattersley AT, Bowden DW, Collins FS, Atzmon G, Chambers JC, Spector TD, Laakso M, Strom TM, Bell GI, Blangero J, Duggirala R, Tai ES, McVean G, Hanis CL, Wilson JG, Seielstad M, Frayling TM, Meigs JB, Cox NJ, Sladek R, Lander ES, Gabriel S, Burt NP, Mohlke KL, Meitinger T, Groop L, Abecasis G, Florez JC, Scott LJ, Morris AP, Kang HM, Boehnke M, Altshuler D, McCarthy MI (2016) The genetic architecture of type 2 diabetes. *Nature* 536:41–47
51. Watanabe RM, Allayee H, Xiang AH, Trigo E, Hartiala J, Lawrence JM, Buchanan TA (2007) Transcription factor 7-like 2 (TCF7L2) is associated with gestational diabetes mellitus and interacts with adiposity to alter insulin secretion in Mexican Americans. *Diabetes* 56:1481–1485
52. Kjos SL, Peters RK, Xiang A, Henry OA, Montoro M, Buchanan TA (1995) Predicting future diabetes in Latino women with gestational diabetes. *Diabetes* 44:586–591
53. Kim C, Newton KM, Knopp RH (2002) Gestational diabetes and the incidence of type 2 diabetes. *Diabetes Care* 25:1862–1868
54. Boston RC, Stefanovski D, Moate PJ, Sumner AE, Watanabe RM, Bergman RN (2003) MINMOD millennium: a computer program to calculate glucose effectiveness and insulin sensitivity from the frequently sampled intravenous glucose tolerance test. *Diabetes Technol Ther* 5:1003–1015
55. Hücking K, Watanabe RM, Stefanovski D, Bergman RN (2008) OGTT-derived measures of insulin sensitivity are confounded by factors other than insulin sensitivity itself. *Obesity* 16:1938–1945
56. Grarup N, Sandholt CH, Hansen T, Pedersen O (2014) Genetic susceptibility to type 2 diabetes and obesity: from genome-wide association studies to rare variants and beyond. *Diabetologia* 57:1528–1541
57. Deeb SS, Fajas L, Nemoto M, Pihlajamäki J, Mykkänen L, Kuusisto J, Laakso M, Fujimoto W, Auwerx J (1998) A Pro12Ala substitution in PPAR γ 2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. *Nat Genet* 20:284–287
58. Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, Tuomi T, Gaudet D, Hudson TJ, Daly M, Groop L, Lander ES (2000) The common PPAR γ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* 26:76–80
59. Ingelsson E, Langenberg C, Hivert MF, Prokopenko I, Lyssenko V, Dupuis J, Magi R, Sharp S, Jackson AU, Assimes TL, Shrader P, Knowles JW, Zethelius B, Abbasi FA, Bergman RN, Bergmann A, Berne C, Boehnke M, Bonnycastle LL, Bornstein SR, Buchanan TA, Bumpstead SJ, Bottcher Y, Chines P, Collins FS, Cooper CC, Dennison EM, Erdos MR, Ferrannini E, Fox CS, Graessler J, Hao K, Isomaa B, Jameson KA, Kovacs P, Kuusisto J, Laakso M, Ladenvall C, Mohlke KL, Morken MA, Narisu N, Nathan DM, Pascoe L, Payne F, Petrie JR, Sayer AA, Schwarz PE, Scott LJ, Stringham HM, Stumvoll M, Swift AJ, Syvanen AC, Tuomi T, Tuomilehto J, Tonjes A, Valle

- TT, Williams GH, Lind L, Barroso I, Quertermous T, Walker M, Wareham NJ, Meigs JB, McCarthy MI, Groop L, Watanabe RM, Florez JC (2010) Detailed physiologic characterization reveals diverse mechanisms for novel genetic loci regulating glucose and insulin metabolism in humans. *Diabetes* 59:1266–1275
60. Dimas AS, Lagou V, Barker A, Knowles JW, Magi R, Hivert MF, Benazzo A, Rybin D, Jackson AU, Stringham HM, Song C, Fischer-Rosinsky A, Boesgaard TW, Grarup N, Abbasi FA, Assimes TL, Hao K, Yang X, Lecocour C, Barroso I, Bonnycastle LL, Bottcher Y, Bumpstead S, Chines PS, Erdos MR, Graessler J, Kovacs P, Morken MA, Narisu N, Payne F, Stancakova A, Swift AJ, Tonjes A, Bornstein SR, Cauchi S, Froguel P, Meyre D, Schwarz PE, Haring HU, Smith U, Boehnke M, Bergman RN, Collins FS, Mohlke KL, Tuomilehto J, Quertermous T, Lind L, Hansen T, Pedersen O, Walker M, Pfeiffer AF, Spranger J, Stumvoll M, Meigs JB, Wareham NJ, Kuusisto J, Laakso M, Langenberg C, Dupuis J, Watanabe RM, Florez JC, Ingelsson E, McCarthy MI, Prokopenko I (2013) Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes* 63:2158–2171
 61. Ader M, Stefanovski D, Richey JM, Kim SP, Kolka CM, Ionut V, Kabir M, Bergman RN (2014) Failure of homeostatic model assessment of insulin resistance to detect marked diet-induced insulin resistance in dogs. *Diabetes* 63:1914–1919
 62. Xiang AH, Watanabe RM, Buchanan TA (2014) HOMA and Matsuda indices of insulin sensitivity: poor correlation with minimal model-based estimates of insulin sensitivity in longitudinal settings. *Diabetologia* 57:334–338
 63. Buchanan TA, Watanabe RM, Xiang AH (2010) Limitations in surrogate measures of insulin resistance. *J Clin Endocrinol Metab* 95:4874–4876
 64. Rasmussen-Torvik LJ, Pankow JS, Jacobs DR, Steffen LM, Moran AM, Steinberger J, Sinaiko AR (2007) Heritability and genetic correlations of insulin sensitivity measured by the euglycaemic clamp. *Diabetic Med* 24:1286–1289
 65. Bergman RN, Zaccaro DJ, Watanabe RM, Haffner SM, Saad MF, Norris JM, Wagenknecht LE, Hokason JE, Rotter JI, Rich SS (2003) Minimal model-based insulin sensitivity has greater heritability and a different genetic basis than homeostasis model assessment or fasting insulin. *Diabetes* 52:2168–2174
 66. Pacini G, Bergman RN (1986) MINMOD: a computer program to calculate insulin sensitivity and pancreatic responsiveness from the frequently sampled intravenous glucose tolerance test. *Comput Methods Prog Biomed* 23:113–122
 67. Olefsky J, Farquhar JW, Reaven G (1973) Relationship between fasting plasma insulin level and resistance to insulin-mediated glucose uptake in normal and diabetic subjects. *Diabetes* 22:507–513
 68. Andres R, Swerdloff R, Pozefsky T, Coleman D (1966) Manual feedback technique for the control of blood glucose concentration. In: Skeegs LJ (ed) *Automation in analytical chemistry*. Mediad, Inc., New York, pp 486–491
 69. Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC (1985) Homeostasis model assessment: insulin resistance and β -cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* 28:412–419
 70. Stumvoll M, Mitrakou A, Pimenta W, Jenssen T, Yki-Järvinen H, Van Haefeten T, Renn W, Gerich J (2000) Use of the oral glucose tolerance test to assess insulin release and insulin sensitivity. *Diabetes Care* 23:295–301
 71. Belfiore F, Iannello S, Volpicelli G (1998) Insulin sensitivity indices calculated from basal and OGTT-induced insulin, glucose, and FFA levels. *Mol Genet Metab* 63:134–141
 72. Matsuda M, DeFronzo RA (1999) Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp. *Diabetes Care* 22:1462–1470
 73. Gutt M, Davis CL, Spitzer SB, Llabre MM, Kumar M, Czarnecki EM, Schneiderman N, Skyler JS, Marks JB (2000) Validation of the insulin sensitivity index (ISI(0,120)): comparison with other measures. *Diabetes Res Clin Pract* 47:177–184
 74. Ren J, Xiang AH, Trigo E, Takayanagi M, Beale E, Lawrence JM, Hartiala J, Richey JM, Allayee H, Buchanan TA, Watanabe RM (2014) Genetic variation in MTNR1B is associated with gestational diabetes mellitus and contributes only to the absolute level of beta cell compensation in Mexican Americans. *Diabetologia* 57:1391–1399
 75. Li C, Qiao B, Zhan Y, Peng W, Chen ZJ, Sun L, Zhang J, Zhao L, Gao Q (2013) Association between genetic variations in MTNR1A and MTNR1B genes and gestational diabetes mellitus in Han Chinese women. *Gynecol Obstet Investig* 76:221–227

76. Nagorny CL, Sathanoori R, Voss U, Mulder H, Wierup N (2011) Distribution of melatonin receptors in murine pancreatic islets. *J Pineal Res* 50:412–417
77. Wang L, Wang Y, Zhang X, Shi J, Wang M, Wei Z, Zhao A, Li B, Zhao X, Xing Q, He L (2010) Common genetic variation in *MTNR1B* is associated with serum testosterone, glucose tolerance, and insulin secretion in polycystic ovary syndrome patients. *Fertil Steril* 94(2486–2489):2489
78. Florez JC (2008) Newly identified loci highlight beta cell dysfunction as a key cause of type 2 diabetes: where are the insulin resistance genes? *Diabetologia* 51:1100–1110
79. Watanabe RM (2010) The genetics of insulin resistance: where's Waldo? *Curr Diab Rep* 10:476–484
80. Pearson ER, Flechtner I, Njolstad PR, Malecki MT, Flanagan SE, Larkin B, Ashcroft FM, Klimes I, Codner E, Iotova V, Slingerland AS, Shield J, Robert JJ, Holst JJ, Clark PM, Ellard S, Sovik O, Polak M, Hattersley AT (2006) Switching from insulin to oral sulfonylureas in patients with diabetes due to *Kir6.2* mutations. *N Engl J Med* 355:467–477
81. Slingerland AS, Nuboer R, Hadders-Algra M, Hattersley AT, Bruining GJ (2006) Improved motor development and good long-term glycaemic control with sulfonylurea treatment in a patient with the syndrome of intermediate developmental delay, early-onset generalised epilepsy and neonatal diabetes associated with the V59M mutation in the *KCNJ11* gene. *Diabetologia* 49:2559–2563
82. Palmer ND, Wagenknecht LE, Langefeld CD, Wang N, Buchanan TA, Xiang AH, Allayee H, Bergman RN, Raffel LJ, Chen YD, Haritunian T, Fingerlin T, Goodarzi MO, Taylor KD, Rotter JI, Watanabe RM, Bowden DW (2016) Improved performance of dynamic measures of insulin response over surrogate indices to identify genetic contributors of type 2 diabetes: the GUARDIAN consortium. *Diabetes* 65:2072–2080
83. Black MH, Fingerlin TE, Allayee H, Zhang W, Xiang AH, Trigo E, Hartiala J, Lehtinen AB, Haffner SM, Bergman RN, McEachin RC, Kjos SL, Lawrence JM, Buchanan TA, Watanabe RM (2008) Evidence of interaction between peroxisome proliferator-activated receptor- γ 2 and hepatocyte nuclear factor-4 α contributing to variation in insulin sensitivity in Mexican Americans. *Diabetes* 57:1048–1056
84. Li X, Allayee H, Xiang AH, Trigo E, Hartiala J, Lawrence JM, Buchanan TA, Watanabe RM (2009) Variation in *IGF2BP2* interacts with adiposity to alter insulin sensitivity in Mexican Americans. *Obesity* 17:729–736
85. Li X, Shu YH, Xiang AH, Trigo E, Kuusisto J, Hartiala J, Swift AJ, Kawakubo M, Stringham HM, Bonnycastle LL, Lawrence JM, Laakso M, Allayee H, Buchanan TA, Watanabe RM (2009) Additive effects of genetic variation in *GCK* and *G6PC2* on insulin secretion and fasting glucose. *Diabetes* 58:2946–2953
86. Shu YH, Hartiala J, Xiang AH, Trigo E, Lawrence JM, Allayee H, Buchanan TA, Bottini N, Watanabe RM (2009) Evidence for sex-specific associations between variation in acid phosphatase locus 1 (*ACP1*) and insulin sensitivity in Mexican Americans. *J Clin Endocrinol Metab* 94:4094–4102
87. Kang ES, Park SY, Kim HJ, Ahn CW, Nam M, Cha BS, Lim SK, Kim KR, Lee HC (2005) The influence of adiponectin gene polymorphism on the rosiglitazone response in patients with type 2 diabetes. *Diabetes Care* 28:1139–1144
88. Sesti G, Laratta E, Cardellini M, Andreozzi F, Del Guerra S, Irace C, Gnasso A, Grupillo M, Lauro R, Hribal ML, Perticone F, Marchetti P (2006) The E23K variant of *KCNJ11* encoding the pancreatic beta-cell adenosine 5'-triphosphate-sensitive potassium channel *Kir6.2* is associated with an increased risk of secondary failure to sulfonylurea in patients with type 2 diabetes. *J Clin Endocrinol Metab* 91:2334–2338
89. Sun H, Gong ZC, Yin JY, Liu HL, Liu YZ, Guo ZW, Zhou HH, Wu J, Liu ZQ (2008) The association of adiponectin allele 45T/G and -11377C/G polymorphisms with type 2 diabetes and rosiglitazone response in Chinese patients. *Br J Clin Pharmacol* 65:917–926
90. Holstein A, Hahn M, Stumvoll M, Kovacs P (2009) The E23K variant of *KCNJ11* and the risk for severe sulfonylurea-induced hypoglycemia in patients with type 2 diabetes. *Horm Metab Res* 41:387–390
91. Wolford JK, Yeatts KA, Dhanjal SK, Black MH, Xiang AH, Buchanan TA, Watanabe RM (2005) Sequence variation in *PPARG* may underlie differential response to troglitazone. *Diabetes* 54:3319–3325

Chapter 19

Identification of Genes for Hereditary Hemochromatosis

Glenn S. Gerhard, Barbara V. Paynton, and Johanna K. DiStefano

Abstract

Hereditary hemochromatosis (HH) is one of the most common genetically transmitted conditions in individuals of Northern European ancestry. The disease is characterized by excessive intestinal absorption of dietary iron, resulting in pathologically high iron storage in tissues and organs. If left untreated, HH can damage joints and organs, and eventually lead to death. There are four main classes of HH, as well as five individual molecular subtypes, caused by mutations in five genes, and the approaches implemented in the discovery of each HH type have specific histories and unique aspects. In this chapter, we review the genetics of the different HH types, including the strategies used to detect the causal variants in each case and the manner in which genetic variants were found to affect iron metabolism.

Key words Hemochromatosis, Linkage mapping, Iron absorption, Iron overload, HFE, Hemojuvelin (HJV), Hepcidin (HAMP), Transferrin receptor 2 (TFR2), Ferroportin 1 (SLC40A1)

1 The Importance of Iron Metabolism

Iron plays a central role in biology. It is required by almost all living organisms, and its wide range of reduction–oxidation states, from -2 to $+6$, which enables electron transfer reactions, allows iron to serve as the key oxygen carrier in eukaryotes. The elemental form of iron functions as a cofactor for a variety of proteins, such as the iron–sulfur cluster proteins, while the heme form of iron provides a prosthetic group for a variety of proteins involved in critical biological processes, such as oxygen transport, electron transfer, and metabolism of xenobiotics. Iron must be obtained from the environment for most species, including humans, who derive it exclusively from the diet under normal circumstances.

Dietary iron is absorbed by the intestinal enterocyte as either nonheme (inorganic) iron or heme (protoheme IX) iron [1]. Dietary heme iron is taken up as an intact metalloporphyrin molecule and has much greater bioavailability than inorganic iron [2]. Relative to nonheme iron [3], the molecular pathways involved in the intestinal uptake of heme iron are still not well defined [4–6].

Abnormalities in iron metabolism are significant medical problems. Iron deficiency is estimated to affect more than 30% of the global population, or about two billion people [7]. Iron deficiency commonly occurs in women during pregnancy because of the need for iron by the developing fetus, and is associated with poor maternal and fetal outcomes. In Africa and Asia, approximately 50% of children <5 years of age manifest iron deficiency anemia [8]. In the United States, the prevalence of iron storage deficiencies, at levels sufficient to cause anemia, in infants and toddlers has been estimated at ~1–2%, and ~3–4% in children from lower income families [9]. These data are consistent with the reduced incidence of iron deficiency in countries whose populations eat meat, whereas iron deficiency is prevalent in geographic areas where equivalent amounts of dietary iron are found in grains and vegetables [10]. Epidemiological studies indicate that dietary iron bioavailability influences body iron stores over time, with heme iron being a more important factor than nonheme iron [11]. Approximately two-thirds of body iron in Western populations is derived from heme iron, although heme iron constitutes only about one third of dietary iron [12].

In an evolutionary context, the dramatically reduced dietary availability of iron that accompanied the transition from a hunting/gathering-based to an agriculture-based society corresponded with an increased risk of iron deficiency, and likely led to the selection of genetic forms of iron overload [13], i.e., hereditary hemochromatosis [14].

2 HH as a Genetic Disease

In 1865, the French physician, Armand Trousseau, provided the first report of a syndrome with findings of cirrhosis, diabetes, and cutaneous hyperpigmentation [15], which was followed by a demonstration of iron pigment-staining of tissues in 1871 [16]. Hanot suggested the term “diabète bronze” and the liver “cirrhose pigmentaire diabetique” in 1886, while 3 years later, the disorder was referred to as “hämochromatose,” indicating that the iron-containing pigment in tissues came from the blood [17].

In 1935, Sheldon published an analysis of previous reports of HH and concluded that “the fact of an occasional familial incidence must obviously be taken into account in any theory regarding the origin of the disease” [18, 19]. But it was not until 1975 that a major breakthrough occurred, when Marcel Simon published that “(1) the association of the disease with hemochromatosis allele maps closer to locus A than to locus B; (2) B antigens B7 and B14 are not independent markers of the hemochromatosis allele; (3) the HLA marking of the hemochromatosis allele is haplotypic; (4) the present geographic distribution pattern of this marking

could have resulted from a unique mutation followed by chromosome recombinations and population migrations; however, (5) no formal proof of this probable hypothesis can be given.” [20]. The last point, likely a spurious demand of a reviewer, has since been borne out of whole genome sequencing of DNA obtained from ancient humans that found HH to be the first Mendelian disease variant identified in prehistory [21].

3 Hereditary Hemochromatosis (HH)

HH is characterized by excessive intestinal absorption of dietary iron, leading to a pathological increase of iron stores in tissues and organs. If left untreated, HH can produce substantial damage to the liver, pancreas, heart, and joints. This damage eventually leads to severe complications, including cirrhosis and liver cancer, diabetes, heart disease, arthritis, neurodegeneration, and when left untreated, hemochromatosis can be fatal.

The etiology of HH is genetic and follows different modes of inheritance, as discussed fully in the following sections. Caucasians of Northern European ancestry are at the highest risk for developing type 1 HH; one in every 200 individuals is estimated to have the disease [22]. However, other genetic variants lead to hemochromatosis, independent of ethnicity. In the USA, approximately 7% of Caucasians and 4% of African Americans have dysregulated iron metabolism, and are at risk for developing complications related to iron overload [23].

Because the regulation of iron metabolism is still not well defined (4–6), the mechanisms by which hemochromatosis develops are not clearly understood. Our understanding of the pathophysiology of the disease became clearer through the identification of the HFE gene (discussed below), where a genetic mutation was found to dramatically increase levels of iron absorption. Under normal conditions, HFE facilitates the binding of transferrin, the primary carrier protein of iron in the blood. When iron levels become depleted, transferrin concentrations rise, and when transferrin levels are high, HFE increases the rate of intestinal iron release. Thus, when the ability of HFE to bind transferrin becomes compromised, as occurs in type 1 HH, levels of transferrin increase, and the intestines release iron as if the body were iron-deficient. Eventually, the excess iron leads to overload storage in the tissues, leading to symptoms of hemochromatosis. However, because HH patients with HFE mutations have variable manifestations of the disease, there are clearly multiple factors that regulate iron metabolism and storage.

Table 1
Classification of hereditary hemochromatosis

Type	Gene/genotype
1a	HFE Cys282Tyr homozygosity
1b	HFE Cys282Tyr/His63Asp compound heterozygosity
1c	Other HFE genotypes, Ser65Cys, etc.
2a	Hemojuvelin (HJV)
2b	Hepcidin (HAMP)
3	Transferrin receptor 2 (TFR2)
4	Ferroportin 1 (SLC40A1)

4 Genetic Characterization of HH Types

HH can be classified as type 1 or HFE-related, with several molecular subtypes, and non-HFE-related HH types 2, 3, and 4 (Table 1). The approaches used to identify the genes for each of the types of HH have their own history and unique aspects. In the following section, we describe the genetic basis of the different types of HH and the manner in which the dysfunction of the genes affect iron metabolism to cause iron overload.

5 Types of HH

5.1 HH Type 1

The key observation for the initial mapping of the locus responsible for HH type 1 was the association of the disease with HLA-A3 of the major histocompatibility region on chromosome 6p. Subsequent work localized the gene to within 1–2 centimorgans (cM) of HLA-A [24]. Linkage disequilibrium studies were consistent with a strong founder effect, initially suggested by the HLA-A3 association. Narrowing the locus, however, proved to be challenging due to contradictory findings. For example, some studies of pedigrees with recombinant chromosomes were consistent with a chromosomal location centromeric to HLA-F, while others suggested a telomeric position. Additional linkage disequilibrium data refined the region to within a megabase or more telomeric to the MHC. However, most of the data supporting this localization suffered from a number of drawbacks, including low marker density, lack of informative recombinant individuals in pedigrees, and extended regions of linkage disequilibrium within the MHC, all of which hampered accurate locus refinement. Given these issues, Mercator Genetics, a private biotech company, joined

the international search for the gene. They used results from several groups to focus on a region telomeric to the MHC that ultimately resulted in the identification of the HH gene [25].

The initial strategy used by Mercator Genetics was to assemble an overlapping set of yeast artificial chromosome (YACs) from the HLA-A gene to a marker located more than 8 Mb away. They constructed a map of sequence tagged sites, which are unique DNA sequences easily detected by PCR, of the overlapping YAC contigs spanning the region. Using these resources, a set of short tandem repeat polymorphic markers and single base-pair substitution markers was generated and used to genotype 101 patients with a proven diagnosis of HH and 64 unaffected individuals. At each marker, the allele that was found at a higher frequency in the HH patients relative to controls was considered to be part of an ancestral haplotype. A measure of linkage disequilibrium was calculated for each marker to define a region of approximately 600 kb showing the strongest evidence of linkage. An estimate of excess homozygotes was also performed to support the linkage data.

A multistage haplotype analysis of 46 HH chromosomes (chromosome 6) that had been isolated in somatic cell hybrids was performed to perform unambiguous phasing. With this approach, a region of about 400 kb was identified as an ancestral region. In a second stage, additional chromosomes were aligned, further narrowing the candidate region shared by affected individuals to about 250 kb.

Within this region, candidate genes were identified using several approaches. One was direct cDNA selection, in which a library of cDNAs was hybridized to immobilized genomic clones from the critical region, which were then eluted, amplified, and cloned [26]. A second approach utilized exon trapping, where genomic DNA fragments were inserted into an intron of the HIV-1 tat gene and transcribed from a vector containing the SV40 early promoter following transfection into COS cells [27]. If an exon with intronic flanking sequences was present, it would be retained in the mature polyA⁺ RNA. The third approach employed de novo sequencing, which identified three novel genes in the critical region [25]. Subsequent sequencing of the entire 250 kb region, considered a monumental task at the time, resulted in the identification of 15 genes within the critical region.

In this study, the sequences of two patients homozygous for the ancestral haplotype were compared with those from two unaffected controls. Two of the 15 genes contained sequences with variants that predicted amino acid changes. The only consistent ancestral variant was a C->A at nucleotide 845 of a HLA- gene, predicted to result in a cysteine to tyrosine substitution at position 282 (C282Y) of the protein. The substitution was postulated to abrogate a disulfide bond in the α 3 domain. An oligonucleotide-

ligation assay was used to analyze DNA from 178 HH patients, in whom 148 were homozygous for the C282Y variant. Further sequencing was performed in nine patients heterozygous for C282Y, identifying a C->G variant in exon 2 that predicted a histidine to aspartic acid mutation at position 63 (H63D). The HH gene was designated HLA-H due to its homology to HLA class I genes, although the name had been published previously, designating a likely pseudogene [28]. The WHO Nomenclature Committee for Factors of the HLA System and the HuGO Genome Nomenclature Committee approved the designation HFE for High FE (iron) [24].

HFE variants identified since the identification of C282Y and H63D include a 56,097 bp deletion of the entire HFE gene and 3 histone-encoding genes, a 32,744 bp deletion of the entire HFE gene, a c.-20G>A variant in the 5'UTR that creates a new ATG start codon, p.Gly43Asp, p.Leu46Trp, a 22 bp deletion resulting in p.Leu50Cysfs*31, p.Ser65Cys, p.Arg66Cys, p.Arg67Cys, p.Val68-Glyfs*20, p.Arg71*, p.Gly93Arg, a single bp deletion resulting in p.Trp94Glyfs*117, p.Ile105Thr, p.Glu114Lys, p.Tyr138*, a single bp deletion resulting in p.Ala158Glnfs*53, a single bp deletion resulting in p.Arg161Glyfs*50, p.Glu168Gln, p.Glu168*, p.Trp169*, p.Leu183Pro, c.616+1G>T intronic splicing variant causing skipping of exon 3, p.Arg224Gln, p.Tyr230Phe, a 3 bp deletion resulting in p.Tyr231del, p.Gln233*, p.Val256Ile, a single base duplication in p.Trp267Leufs*80, p.Cys282Ser, p.Gln283*, p.Gln283Pro, p.Val295Glu, p.Arg330Met, c.1006+1G>A intronic splicing variant causing skipping of exon 5, and a 13 bp deletion resulting in p.His341Leufs*119 [29].

5.2 HH Type 2a

“Juvenile” or Type 2 hemochromatosis is clinically and phenotypically more severe than Type 1 HFE hemochromatosis [30]. In contrast to HFE HH, which disproportionately affects males, HH Type 2a has equal distribution between the sexes. The phenotype presents with liver enlargement and pain in childhood, hypogonadotropism and joint problems during the teenage years, and cirrhosis, cardiac arrhythmias, and heart failure by the third decade of life.

The approach used to identify the Type 2a “Juvenile” hemochromatosis gene was similar to that used to associate HFE with the HLA locus [31]. First, mapping studies were used to locate a chromosomal region segregating with disease. A linkage analysis of 375 microsatellite markers in 12 patients and 27 unaffected family members yielded a map with a resolution of ~10 cM. Radiation hybrid mapping was then performed with selected microsatellite markers or sequence-tagged sites using a commercial hybrid panel. The genome assembly at the time of analysis (April 2003, build 33) was incomplete with duplicated regions and multiple gaps. Despite this handicap, the investigators were able to define the boundary of linkage with available recombinants.

The strongest evidence for linkage (logarithm of the odds [LOD] score >5.0) was found with markers on chromosome 1. The same group of investigators leveraged this finding to conduct a larger study of 12 unrelated families affected by juvenile hemochromatosis from Canada, France, and Greece [32]. The hepcidin gene was sequenced in all 12 families to exclude their classification as HH Type 2b. Verification of linkage with chromosome 1q using the same initial marker set was performed along with additional microsatellite markers. The Greek families were particularly informative with 9/10 showing a large tract of extended homozygosity in the region of the linkage peak on chromosome 1q. A total of five different haplotypes in the linked region were found to segregate with disease in the Greek families.

This effort paved the way for positional cloning to identify candidate genes in the linkage region. A total of 21 annotated genes were identified in the critical region. The predicted coding regions for all 21 genes were sequenced, and six rare variants were identified in a previously uncharacterized predicted transcript, LOC148738. One variant was predicted to cause a premature termination codon, one a frameshift mutation, and four were missense variants at evolutionarily conserved residues. The mutations segregated with the individuals affected with juvenile hemochromatosis in a recessive inheritance pattern with complete penetrance. One missense variant, a G320V substitution, was present in over a third of the families. Initially designated HJV, due its lack of homology to HFE, the gene was later named HFE2 [33].

Subsequent studies identified additional HH-causing variants in HFE2, including a single base intronic duplication resulting in c.-89-4dupT that may affect splicing, p.Gln6His, p.Thr19Ala, a single base deletion resulting in p.Leu28Serfs*24, p.Arg54*, p.Gly66*, a three base insertion resulting in p.Gly69dup, p.Arg70Trp, a single base deletion resulting in p.Val74Trpfs*40, p.Cys80Arg, p.Cys80Tyr, p.Ser85Pro, p.Cys89Arg, p.Gly99Val, p.Gly99Arg, p.Leu101Pro, a single base deletion resulting in p.Phe103Serfs*11, p.Ser105Leu, p.Gln116*, p.Cys119Phe, a 12 base deletion resulting in p.Arg131Phefs*111, p.Leu135Arg, a single base deletion resulting in p.Asp149Thrfs*97, p.Leu165*, p.Alal68Asp, p.Phe170Ser, p.Asp172Glu, p.Arg176Cys, p.Trp191Cys, p.Pro192Leu, p.Leu194Pro, p.Asn196Lys, p.Ser205Arg, p.Ile222Asn, a three base deletion resulting in p.Lys234del, p.Asp249His, p.Gly250Val, p.Ser264Leu, a single base insertion resulting in p.Asn269Lysfs*43, p.Ile281Thr, p.Arg288Trp, p.Glu302Lys, p.Ala310Gly, p.Gln312*, a single base insertion resulting in p.Cys321Valfs*21, p.Gly320Val, p.Cys321Trp, a two base deletion-insertion mutation resulting in p.Cys321*, p.Arg326*, a 4 base deletion resulting in p.Ser328Aspfs*10, p.Arg335Gln, a single base deletion resulting in p.Ala343Profs*24, a single base deletion resulting in p.Cys361Valfs*6, p.Asn372Asp, and p.Arg385* [29].

5.3 *HH Type 2b*

The identification of the second juvenile HH locus using established diagnostic criteria [34] did not rely on linkage analysis, likely because of an insufficient number of families. Instead, two pedigrees in which the disease did not segregate with the juvenile HH chromosome 1q locus were used to interrogate the hepcidin locus as a candidate gene. Hepcidin antimicrobial peptide (HAMP) had been implicated in iron metabolism in animal models [35] and was selected as a candidate gene. However, rather than directly sequencing the gene, microsatellite markers were first used to identify a shared span of homozygosity corresponding to a 2.7 cM region on chromosome 19q13 in two probands from one family. Sequencing of the hepcidin exon–intron boundaries, coding exons, and 5' and 3' untranslated regions of affected individuals of the two families identified two mutations.

One mutation was a homozygous frameshift variant, resulting from a single base deletion in exon 2 at position 93 (93delG) and absent in 50 unaffected controls. The other mutation produced a cytosine to thymidine substitution at position 166 in exon 3, creating a stop codon (R56X). Restriction fragment length polymorphism analysis was used to demonstrate segregation within the affected family and its absence in 50 unaffected individuals.

Other HAMP variants associated with HH type 2b include c. -153C>T and c. -72C>T promoter variants, a c. -25G>A single base change in the 5' UTR that creates a new ATG, a single base deletion resulting in p.Gly32Aspfs*88, a 2 base deletion resulting in p.Arg42Serfs*78, a 4 base deletion that results in p.Met50fs, p.Arg56*, p.Arg59Gly, p.Cys70Arg, p.Gly71Asp, p.Arg75*, p.Cys78Tyr, and p.Lys83Arg [29].

5.4 *HH Type 3*

An approach similar to that used for the identification of the Type 2b HH gene was used to identify the Type 3 gene. Six patients from two families with HH and lacking HFE mutations were studied [36]. The patients were members of two unrelated Sicilian families, including a large, inbred family [37]. The consanguinity in this kindred was exploited to map regions of homozygosity using polymorphic markers genotyped in all affected individuals. Pairwise linkage analysis identified a region on chromosome 7q with the highest lod score 4.09. This interval was then confirmed in the second family and further refined to a region <1 cM. A plausible candidate gene, transferrin receptor 2 (TFR2), which has about 2/3 shared sequence identity with TFR1, had recently been localized to chromosome 7q22 using radiation hybrid mapping [38]. Two polymorphic repeats located within the TFR2 locus were identified through analysis of extant sequence data, and used to determine that the TFR2 gene locus was also homozygous in the affected individuals.

The coding region and intron/exon boundaries of the TFR2 gene were subsequently sequenced to identify potential mutations.

A stop-gain variant (C->G), in which a TAC codon for tyrosine was replaced with a TAG stop codon at position 750 in exon 6 of the cDNA (Y250X), was identified. A PCR RFLP assay, based on the generation of a MaeI site by the mutation, was then used to determine segregation in one of the families. The Y250X variant was found to be homozygous in all affected members and heterozygous in all obligate carriers. The mutation was not observed in 50 unaffected controls or in a dozen HH patients lacking HFE mutations.

Interestingly, to scan other patients for the Y250X mutation, an alternative assay was designed because MaeI and MaeI isoschizomers are relatively unstable and present several practical difficulties for higher volume assays [39]. The alternative assay created a restriction site for RsaI by modifying a forward PCR primer to introduce three of the four bases coding the RsaI recognition site at the 3' end of the primer. The Y250X mutation is located at the first base 3' from the forward primer that completes the RsaI site, whereas the wild-type sequence disrupts the site, and prevents cleavage. Using this assay, no Y250X mutations were detected in 63 French HH patients [39]. Subsequent studies have identified additional HH variants in TFR2 including p.Val22Ile, a single base duplication resulting in p.Arg30Profs*31, p.His33Asn, a 41 bp deletion that results in p.Leu85_Ala96delinsPro and deletes a splice site and part of an intron, p.Leu99Val, p.Arg105*, p.Met172Lys, c.614+4A>G intronic splicing variant resulting in skipping of exon 4, c.727-9T>A intronic variant that may affect splicing, p.Tyr250*, p.Val277Leu, p.Phe280Leu, p.Gln306*, p.Gln317*, p.Gly373Asp, p.Ala376Asp, p.Arg396*, p.Asp402Lys, a 2 bp deletion resulting in p.Asn412del, p.Asn412Ile, p.Arg420His, p.Gly430Arg, p.Ala444Thr, p.Ile449Val, p.Arg455Gln, p.Arg468His, p.Leu490Arg, a single base variant resulting in p.Glu491Glu that may affect splicing, p.Tyr504Cys, c.1538-2A>G intronic splicing variant, p.Ile529Asn, a 4 bp deletion and 13 bp insertion resulting in p.Ser531Glnfs*6, c.1606-8A>G intronic splicing variant, a single bp insertion resulting in p.Ser556Alafs*6, p.Val583Ile, a single base duplication resulting in p.Leu615Profs*177, an 11 bp deletion resulting in p.Ala621_Gln624del, p.Gln672*, p.Arg678Pro, p.Gln690Pro, a single bp duplication resulting in p.Met705Hisfs*87, c.2137-1G>A intronic splicing variant, p.Arg730Cys, p.Gly735Ser, p.Thr740Met, p.Ala743Val, p.Arg752His, p.Trp781*, and p.Gly792Arg [29].

5.5 HH Type 4

HH is not always transmitted as an autosomal recessive disorder; some families exhibit dominant inheritance [40, 41]. Two papers published in the same month used similar approaches to identify the gene for Type 4 HH. In one, a genome-wide scan was conducted using 400 short tandem repeat polymorphic Genethon

markers with an average spacing of 10 cM in 96 members of a single family from the Netherlands in whom HH segregated as a dominant trait [42]. However, the HH phenotype was modeled to account for ambiguities in affection status, including age- and sex-related differences in penetrance. Individuals who were considered as possibly affected were classified as unknown disease status. Linkage analysis using an autosomal dominant model with a frequency of 0.001 yielded positive lod scores for several markers on the long arm of chromosome 2 with the highest lod score of 3.01, very close to the simulated estimate for the large pedigree of a maximum lod score of 3.24. Additional markers from the implicated region were genotyped to construct haplotypes to identify regions of shared identity. In addition, recombination events were used to define a critical region of about 9 cM.

With the critical region on chromosome 2q defined, a number of candidate genes were selected based upon information found in public databases and published reports. One gene, SLC11A3, previously designated as FPN1, IREG1, and MTP1 [43] but now named solute carrier family 40 member 1 (SLC40A1), was selected for further investigation based on its role in mediating iron transport from the basolateral surface of intestinal enterocytes to the blood and the presence of an iron-responsive element (IRE) in its 5' untranslated region [44]. Mutation analysis was performed by PCR amplification of exons, >50 bp of flanking intron sequences, and the 5' and 3' UTRs that contained the IRE, followed by direct Sanger sequencing of both strands of the PCR products. A heterozygous A->C missense mutation at position 734 in exon 5, resulting in an aspartic acid-histidine (N144H) substitution, was found in all affected individuals and absent in 200 unaffected family members and healthy Dutch individuals drawn from the same geographical region. Aspartate at the designated position was conserved among vertebrate species and present in a region of shared homology with other divalent metal ion transport proteins.

The second group also studied a family, a large Italian kindred manifesting autosomal dominant HH [45]. They used a panel of 375 polymorphic markers with average spacing of 10 cM and identified a region of linkage on chromosome 2q32. Recombinants enabled localization to a 5 cM region that was then scanned for potential candidate genes. SLC11A3 (now SLC40A1), located in this region, was selected based on its known involvement in iron metabolism. All SLC11A3 exons were amplified and sequenced. Patients with iron overload were found to be heterozygous for a C->A substitution that replaced a highly conserved alanine at amino acid 77 with aspartic acid. The A77D mutation was not found in 25 unaffected family members or in 100 apparently healthy blood Italian donors.

Further characterization of SLC40A1 HH related variants include c.-188A>G 5' UTR variant, a 14 bp 5' UTR deletion

c.-59_-45del, p.Ala45Glu, p.Tyr64His, p.Ala69Thr, p.Ala69Val, p.Ser71Phe, p.Val72Phe, p.Ala77Asp, p.Gly80Ser, p.Gly80Val, p.Arg88Gly, p.Arg88Thr, p.Leu129Pro, p.Asn144His, p.Asn144Asp, p.Asn144Thr, p.Ile152Phe, p.Asp157Asn, p.Asp157Ala, p.Asp157Gly, p.Asp157Tyr, p.Trp158Leu, p.Trp158Cys, a 2 bp deletion resulting in p.Val162del, p.Asn174Ile, p.Arg178Gln, p.Ile180Thr, p.Asp181Val, p.Gln182His, p.Asn185Asp, p.Asn185Thr, p.Gly204Ser, p.Thr230Asn, p.Ala232Asp, p.Leu233Pro, p.Lys240Glu, p.Gln248His, p.Met266Thr, p.Gly267Asp, p.Asp270Val, p.Gly323Val, p.Cys326Ser, p.Cys326Tyr, p.Cys326Phe, p.Ser338Arg, p.Leu345Phe, p.Ile351Val, p.Arg371Trp, p.Arg371Gln, p.Pro443Leu, p.Gly468Ser, p.Arg489Lys, p.Arg489Ser, p.Gly490Ser, p.Gly490Asp, p.Tyr501Cys, p.Asp504Asn, p.His507Arg, and p.Arg561Gly [29].

6 Conclusions

Despite the pleiotropy and complexity of the phenotypic manifestations resulting from iron overload, HH is caused by pathogenic variants in just five genes. The discovery of the two major variants underlying Type I HH in the HFE gene, C282Y and H63D, using linkage and sequencing, presaged the relatively rapid identification of the other four genes related to HH. Identification of the Juvenile HH Type 2a HJV gene and the HH Type 4 SLC40A1 genes used similar approaches while the Juvenile HH Type 2 HAMP gene was identified using a Bayesian candidate gene analysis, as was the HH Type 3 TFR2 gene. All of the genes were found years prior to the modern era of next generation sequencing and relied on relatively few, but informative, pedigrees and careful phenotypic characterization. Despite the increased power of current genetic methods, the successful elucidation of the HH genes using robust phenotyping of related individuals may be a historical lesson in the key strategy for successful gene identification. The willingness of funding bodies and peer reviewers to authorize projects in which more resources can be devoted to cohorts with in-depth and accurate phenotypic data may yield similar results in the hunt for genes for other disorders such as diabetes, obesity, and neurodegeneration.

References

1. Gulec S, Anderson GJ, Collins JF (2014) Mechanistic and regulatory aspects of intestinal iron absorption. *Am J Physiol Gastrointest Liver Physiol* 307(4):G397–G409
2. Hoppe M, Brun B, Larsson MP, Moraes L, Hulthen L (2013) Heme iron-based dietary intervention for improvement of iron status in young women. *Nutrition* 29(1):89–95
3. Silva B, Faustino P (2015) An overview of molecular basis of iron metabolism regulation and the associated pathologies. *Biochim Biophys Acta* 1852(7):1347–1359

4. West AR, Oates PS (2008) Mechanisms of heme iron absorption: current questions and controversies. *World J Gastroenterol* 14 (26):4101–4110
5. Hooda J, Shah A, Zhang L (2014) Heme, an essential nutrient from dietary proteins, critically impacts diverse physiological and pathological processes. *Nutrients* 6(3):1080–1102
6. Fleming MD, Hamza I (2012) Mitochondrial heme: an exit strategy at last. *J Clin Invest* 122 (12):4328–4330
7. Bailey RL, West KP Jr, Black RE (2015) The epidemiology of global micronutrient deficiencies. *Ann Nutr Metab* 66(Suppl 2):22–33
8. Miller JL (2013) Iron deficiency anemia: a common and curable disease. *Cold Spring Harb Perspect Med* 3(7):a011866
9. McDonagh MS, Blazina I, Dana T, Cantor A, Bougatsos C (2015) Screening and routine supplementation for iron deficiency anemia: a systematic review. *Pediatrics* 135(4):723–733
10. Uzel C, Conrad ME (1998) Absorption of heme iron. *Semin Hematol* 35(1):27–34
11. Hunt JR, Roughead ZK (2000) Adaptation of iron absorption in men consuming diets with high or low iron bioavailability. *Am J Clin Nutr* 71(1):94–102
12. Conrad ME, Umbreit JN (2000) Iron absorption and transport—an update. *Am J Hematol* 64(4):287–298
13. Distante S, Robson KJ, Graham-Campbell J, Arnaiz-Villena A, Brissot P, Worwood M (2004) The origin and spread of the HFE-C282Y haemochromatosis mutation. *Hum Genet* 115(4):269–279
14. Felitti VJ, Beutler E (1999) New developments in hereditary hemochromatosis. *Am J Med Sci* 318(4):257–268
15. Trousseau A (1865) Glycosurie, diabète sucre. In: *Clinique médicale de l'Hôtel-Dieu de Paris*, vol 2. J.-B. Balliere, Paris, pp 663–698
16. Troisier M (1871) Diabète sucre. *Bull Soc Anat (Paris)* 44:231–235
17. von Recklinghausen FD (1889) Ueber Hamochromatose. *Bericht der Naturforscherversammlung zu Heidelberg*. p 324
18. Sheldon JH (1935) *Haemochromatosis*. Oxford University Press, H. Milford, Oxford, p 382
19. McKusick VA (1998) *Mendelian inheritance in man: a catalog of human genes and genetic disorders*, vol 1. Johns Hopkins University Press, Baltimore, MD
20. Simon M, Le Mignon L, Fauchet R, Yaouanq J, David V, Edan G, Bourel M (1987) A study of 609 HLA haplotypes marking for the hemochromatosis gene: (1) mapping of the gene near the HLA-A locus and characters required to define a heterozygous population and (2) hypothesis concerning the underlying cause of hemochromatosis-HLA association. *Am J Hum Genet* 41(2):89–105
21. Cassidy LM, Martiniano R, Murphy EM, Teasdale MD, Mallory J, Hartwell B, Bradley DG (2016) Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proc Natl Acad Sci U S A* 113 (2):368–373
22. Merryweather-Clarke AT, Pointon JJ, Shearman JD, Robson KJ (1997) Global prevalence of putative haemochromatosis mutations. *J Med Genet* 34(4):275–278
23. Barton JC, Acton RT, Dawkins FW, Adams PC, Lovato L, Leiendecker-Foster C, McLaren CE, Reboussin DM, Speechley MR, Gordeuk VR, McLaren GD, Sholinsky P, Harris EL (2005) Initial screening transferrin saturation values, serum ferritin concentrations, and HFE genotypes in whites and blacks in the Hemochromatosis and Iron Overload Screening Study. *Genet Test* 9(3):231–241
24. Barton JC, Edwards CQ, Acton RT (2015) HFE gene: structure, function, mutations, and associated iron abnormalities. *Gene* 574 (2):179–192
25. Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, Domingo R Jr, Ellis MC, Fullan A, Hinton LM, Jones NL, Kimmel BE, Kronmal GS, Lauer P, Lee VK, Loeb DB, Mapa FA, McClelland E, Meyer NC, Mintier GA, Moeller N, Moore T, Morikang E, Prass CE, Quintana L, Starnes SM, Schatzman RC, Brunke KJ, Drayna DT, Risch NJ, Bacon BR, Wolff RK (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat Genet* 13(4):399–408
26. Lovett M, Kere J, Hinton LM (1991) Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc Natl Acad Sci U S A* 88(21):9628–9632
27. Church DM, Stotler CJ, Rutter JL, Murrell JR, Trofatter JA, Buckler AJ (1994) Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nat Genet* 6(1):98–105
28. Venditti CP, Harris JM, Geraghty DE, Chorney MJ (1994) Mapping and characterization of non-HLA multigene assemblages in the human MHC class I region. *Genomics* 22 (2):257–266
29. Wallace DF, Subramaniam VN (2016) The global prevalence of HFE and non-HFE hemochromatosis estimated from analysis of next-

- generation sequencing data. *Genet Med* 18 (6):618–626
30. Le Gac G, Ferec C (2005) The molecular genetics of haemochromatosis. *Eur J Hum Genet* 13(11):1172–1185
 31. Roetto A, Totaro A, Cazzola M, Cicilano M, Bosio S, D'Ascola G, Carella M, Zelante L, Kelly AL, Cox TM, Gasparini P, Camaschella C (1999) Juvenile hemochromatosis locus maps to chromosome 1q. *Am J Hum Genet* 64(5):1388–1393
 32. Papanikolaou G, Samuels ME, Ludwig EH, MacDonald ML, Franchini PL, Dube MP, Andres L, MacFarlane J, Sakellaropoulos N, Politou M, Nemeth E, Thompson J, Risler JK, Zaborowska C, Babakaiff R, Radomski CC, Pape TD, Davidas O, Christakis J, Brissot P, Lockitch G, Ganz T, Hayden MR, Goldberg YP (2004) Mutations in HFE2 cause iron overload in chromosome 1q-linked juvenile hemochromatosis. *Nat Genet* 36 (1):77–82
 33. Lee PL, Beutler E, Rao SV, Barton JC (2004) Genetic abnormalities and juvenile hemochromatosis: mutations of the HJV gene encoding hepcidin. *Blood* 103(12):4669–4671
 34. De Gobbi M, Roetto A, Piperno A, Mariani R, Alberti F, Papanikolaou G, Politou M, Lockitch G, Girelli D, Fargion S, Cox TM, Gasparini P, Cazzola M, Camaschella C (2002) Natural history of juvenile haemochromatosis. *Br J Haematol* 117(4):973–979
 35. Nicolas G, Bennoun M, Porteu A, Mativet S, Beaumont C, Grandchamp B, Sirito M, Sawadogo M, Kahn A, Vaulont S (2002) Severe iron deficiency anemia in transgenic mice expressing liver hepcidin. *Proc Natl Acad Sci U S A* 99(7):4596–4601
 36. Camaschella C, Roetto A, Cali A, De Gobbi M, Garozzo G, Carella M, Majorano N, Totaro A, Gasparini P (2000) The gene TFR2 is mutated in a new type of haemochromatosis mapping to 7q22. *Nat Genet* 25(1):14–15
 37. Camaschella C, Fargion S, Sampietro M, Roetto A, Bosio S, Garozzo G, Arosio C, Piperno A (1999) Inherited HFE-unrelated hemochromatosis in Italian families. *Hepatology* 29(5):1563–1564
 38. Kawabata H, Yang R, Hirama T, Vuong PT, Kawano S, Gombart AF, Koeffler HP (1999) Molecular cloning of transferrin receptor 2. A new member of the transferrin receptor-like family. *J Biol Chem* 274(30):20826–20832
 39. Aguilar-Martinez P, Esculie-Coste C, Bismuth M, Giansily-Blaizot M, Larrey D, Schved JF (2001) Transferrin receptor-2 gene and non-C282Y homozygous patients with hemochromatosis. *Blood Cells Mol Dis* 27 (1):290–293
 40. Inheritance of idiopathic haemochromatosis (1977) *Lancet* 1(8021):1106–1107
 41. Pietrangelo A, Montosi G, Totaro A, Garuti C, Conte D, Cassanelli S, Fraquelli M, Sardini C, Vasta F, Gasparini P (1999) Hereditary hemochromatosis in adults without pathogenic mutations in the hemochromatosis gene. *N Engl J Med* 341(10):725–732
 42. Njajou OT, Vaessen N, Joesse M, Berghuis B, van Dongen JW, Breuning MH, Snijders PJ, Rutten WP, Sandkuijl LA, Oostra BA, van Duijn CM, Heutink P (2001) A mutation in SLC11A3 is associated with autosomal dominant hemochromatosis. *Nat Genet* 28 (3):213–214
 43. Haile DJ (2000) Assignment of Slc11a3 to mouse chromosome 1 band 1B and SLC11A3 to human chromosome 2q32 by in situ hybridization. *Cytogenet Cell Genet* 88 (3-4):328–329
 44. Donovan A, Brownlie A, Zhou Y, Shepard J, Pratt SJ, Moynihan J, Paw BH, Drejer A, Barut B, Zapata A, Law TC, Brugnara C, Lux SE, Pinkus GS, Pinkus JL, Kingsley PD, Palis J, Fleming MD, Andrews NC, Zon LI (2000) Positional cloning of zebrafish ferroportin1 identifies a conserved vertebrate iron exporter. *Nature* 403(6771):776–781
 45. Montosi G, Donovan A, Totaro A, Garuti C, Pignatti E, Cassanelli S, Trenor CC, Gasparini P, Andrews NC, Pietrangelo A (2001) Autosomal-dominant hemochromatosis is associated with a mutation in the ferroportin (SLC11A3) gene. *J Clin Invest* 108 (4):619–623

Identification of Driver Mutations in Rare Cancers: The Role of *SMARCA4* in Small Cell Carcinoma of the Ovary, Hypercalcemic Type (SCCOHT)

Jessica D. Lang and William P.D. Hendricks

Abstract

Cancer is a complex genetic disease that can arise through the stepwise accumulation of mutations in oncogenes and tumor suppressor genes in a variety of different tissues. While the varied landscapes of mutations driving common cancer types such as lung, breast, and colorectal cancer have been comprehensively charted, the genetic underpinnings of many rare cancers remain poorly defined. Study of rare cancers faces unique methodological challenges, but collaborative enterprises that incorporate next generation sequencing, reach across disciplines (i.e., pathology, genetic epidemiology, genomics, functional biology, and preclinical modeling), engage advocacy groups, tumor registries, and clinical specialists are adding increasing resolution to the genomic landscapes of rare cancers. Here we describe the approaches and methods used to identify *SMARCA4* mutations, which drive development of the rare ovarian cancer, small cell carcinoma of the ovary, hypercalcemic type (SCCOHT), and point to the broader relevance of this paradigm for future research in rare cancers.

Key words Cancer genetics, Cancer genomics, Genetic epidemiology, Rare cancers, Small cell carcinoma of the ovary, hypercalcemic type (SCCOHT), *SMARCA4*, *Brg1*, *SMARCA2*, *Brm*

1 Introduction

Rare cancers affect <15 out of 100,000 individuals per year [1, 2]. Despite this rarity, such cases comprise the fourth leading cause of death in the United States, and collectively account for 22–27% of cancer burden and 25% of cancer deaths [1–4]. Patient outcomes in rare cancers also lag behind those of common cancers. Five-year overall survival is 47% for rare cancers versus 65% for common cancers, and survival rates have improved more slowly in rare cancers [1, 3]. Many factors likely contribute to poorer outcomes for these patients, including diagnostic and prognostic challenges, lack of effective standardized treatments, limited clinical expertise, and insufficient preclinical or clinical research funding. Clearly, a significant unmet need exists for improved clinical

management of rare cancers built on the foundation of comprehensive molecular characterization. Mapping the genomic basis of rare cancers also has value for a broader understanding of cancer genetics and biology. Rare cancer studies have repeatedly yielded fundamental insights into common cancers. For example, discovery of one of the first tumor suppressor genes, *RBI*, and formulation of the “two-hit hypothesis” of tumor suppressor inactivation both derived from studies of retinoblastoma (diagnosed in 0.35–1.18 per 100,000 individuals per year) [5–8].

Unlike other genetic disorders, cancer can arise from any cell or tissue in the body through mutation of a diverse array of genes governing cell growth and death. The stepwise accumulation of cancer-driving mutations is required for malignant transformation. To this end, cancer gene hunting necessitates integration of family history, clinical and pathologic annotation, and, since the advent of next generation sequencing (NGS), both germline and somatic tumor genomic sequencing in large patient cohorts. However, in the absence of hundreds or even dozens of cases for rare cancers, an even deeper integration of approaches and disciplines is necessary. This chapter highlights how the general challenges presented by studying rare cancers were overcome within the context of the recent identification of the genetic underpinnings of small cell carcinoma of the ovary, hypercalcemic type (SCCOHT) as a case study. While the genetic profile and epidemiology of any given rare cancer will undoubtedly present unique problems, many of the themes described here are conceptually universal.

SCCOHT is a rare form of ovarian cancer affecting young women (mean age at diagnosis: 24 years). Accurate estimates of prevalence and incidence of this disease do not exist due to SCCOHT’s extreme rarity, but around 300 cases have been reported in the literature since its first description in 1979 [9, 10]. Until recently, these tumors were challenging to diagnose. Treatment is aggressive and nonspecific, consisting of high-dose, multi-agent chemotherapy and radiation. Thus, SCCOHT patients face a poor prognosis with a 5-year survival rate of 16% [10]. Prior to the discovery of *SMARCA4* (also known as Brg1) mutations in SCCOHT, its molecular basis was unclear, although it was thought to bear a diploid genome and lack mutations in cancer genes including *BRAF*, *BRCA1/2*, *KRAS*, and *TP53* [11–13]. Remarkably, in a period of time spanning 2013 and 2014, four independent collaborative teams reported inactivating mutation of the SWItch/Sucrose Non-Fermentable (SWI/SNF) complex member and tumor suppressor *SMARCA4* in nearly all SCCOHTs [14–18]. These research teams arrived at a common conclusion using different approaches to identify cancer driver mutations, including targeted Sanger sequencing and NGS (whole genome, exome, and targeted panels) within both affected families and sporadic cases. Subsequent studies concluded that

SMARCA4 mutation and protein loss, concomitant with the loss of the alternative ATPase SMARCA2 (also known as Brm) protein through epigenetic silencing, is pathognomonic for SCCOHT [19, 20]. Here, we discuss the key steps leading to the identification of this driving genetic alteration and we outline the general approach to dissecting the genetic basis of rare cancers (Fig. 1).

2 Methods

2.1 Study Design and Research Consortia

The most obvious challenge facing identification and validation of driver mutations (i.e., those that play a causal role in tumorigenesis) in rare cancers is sample scarcity. Most cancer subtypes are genetically heterogeneous and predominantly sporadic. Large cohorts are required to reach sufficient power to discriminate driver genes from a background of passenger mutations and nonpathogenic normal variation in the general population. However, many rare histologic subtypes of cancer are far more genetically uniform compared to common cancers. For example, pathognomonic mutations in *RBI*, *FOXL2*, or *BRAF* have been discovered in retinoblastoma, granulosa cell tumors, and hairy cell leukemias, respectively [6, 21, 22]. These mutations often occur amidst an otherwise low somatic mutation burden. In these cases, analysis of only a few tumors may lead to discovery of characteristic mutations. Yet, cohort collection remains a challenge both because pathognomonic mutations do not exist in some rare cancer subtypes (e.g., chordomas) [23] and also because even when pathognomonic drivers are discovered, validation studies require additional samples and model systems.

Although study of rare cancers has historically been limited to a small number of academic centers or specialty clinics, cohort collection challenges are now being addressed by advances in global research communication and collaboration. Far-reaching SCCOHT collaborations have relied upon clinical and scientific expertise in genetics, pathology, gynecologic oncology, and epidemiology, in addition to engagement of patients and their families. Research access to high quality, clinically annotated tissue samples requires active clinical partnerships and engagement with specialty clinics to maximize cohort size. Cancer registries such as the International Ovarian and Testicular Stromal Tumor Registry (www.otstregistry.org) are also critical tools for collating rare cancer data and samples utilizing clinician- or patient-driven information collection. Finally, patients with rare cancers are often underserved by traditional funding and public interest in their cancers and thus, tend to be highly motivated to participate in and advocate for research. For example, the Small Cell Ovarian Cancer Foundation (www.smallcellovarian.org) provides advocacy and support services, in addition to managing an SCCOHT registry in partnership with Patient Crossroads.

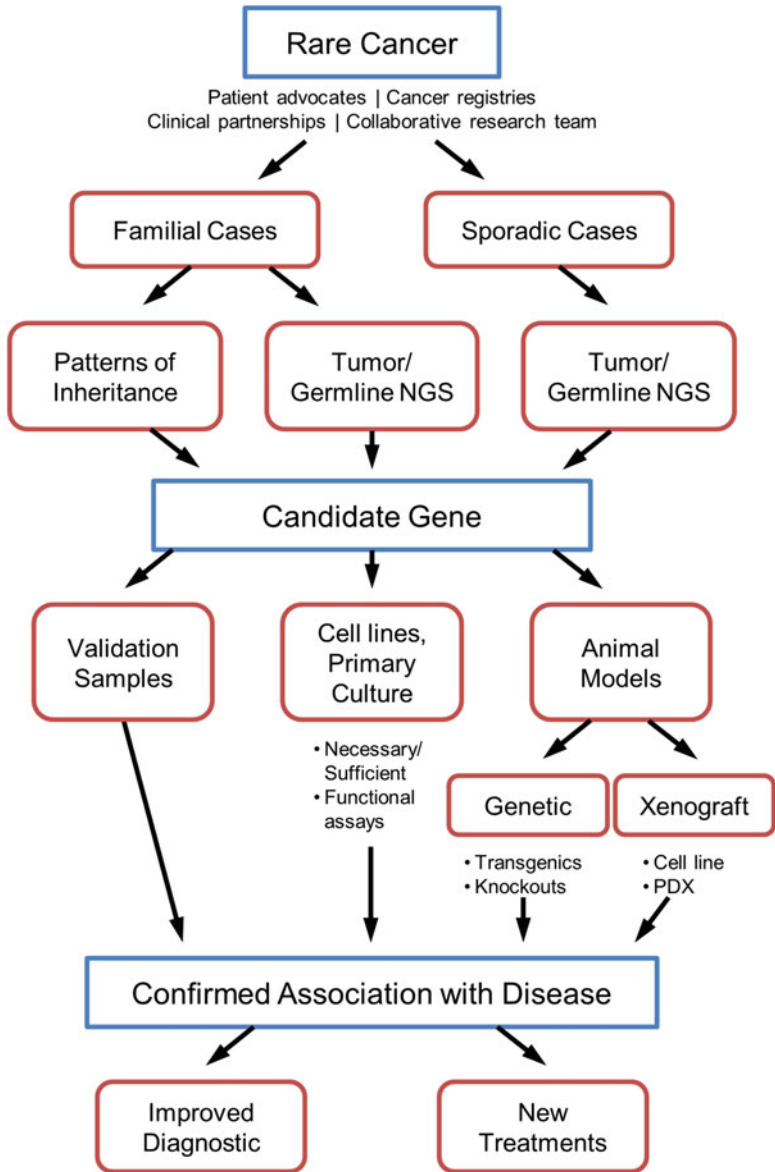


Fig. 1 Generalized approach to cancer gene identification and validation. A flowchart depicts the general process underlying discovery of cancer-driving mutations in rare cancers. This process often begins simply with an unmet need for characterization of a previously understudied histologic tumor subtype, but may also begin with clinical observations of rare cancers that segregate in families. Broad consortia then enable cohort collection and candidate gene (i.e., a gene bearing putative cancer-driving mutations) discovery through genetic epidemiology and next generation sequencing (NGS) of families as well as NGS in sporadic tumors. A causal role for these genes must then be established through validation studies in extended cohorts and prospective functional experiments in cell lines and animal models. A confirmed association of the gene with the disease can then enable diagnostic and therapeutic innovation. *PDX* Patient-derived xenograft

Using analysis of only a few dozen patient tumors and germlines, SCCOHT has been definitively established as a monogenic disorder. Nonetheless, multiple academic, clinical, and advocacy collaborations were critical for rapidly cementing this discovery. As noted, the four groups that identified *SMARCA4* mutations in SCCOHT tumorigenesis combined unique IRB-approved research study designs with clinical pathology, DNA sequencing, and immunohistochemistry (IHC). These landmark studies, whose cohort characteristics are outlined below, together with subsequent follow-up work and case reports, have now identified a total of 96 distinct and predominantly inactivating (coinciding with loss of SMARCA4 protein by IHC) *SMARCA4* mutations in 89 patients [10]:

1. Kupryjańczyk et al. performed *SMARCA4* Sanger sequencing on two SCCOHT patient tumors from the Department of Pathology at the Institute of Oncology in Warsaw, Poland [14].
2. Witkowski et al. performed whole exome tumor and/or germline sequencing in four SCCOHT families and 23 nonfamilial cases or cases with unknown family history from McGill University, Montreal, Quebec, Canada and the Children's Oncology Group, Monrovia, California, USA [18].
3. Jelinic et al. performed targeted NGS of 279 cancer-related genes in 12 tumors and germlines from SCCOHT patients with unknown family history collected at the Memorial Sloan-Kettering Cancer Center, New York, New York, USA, in coordination with other academic centers and patient support groups [15].
4. Ramos et al. performed whole genome, whole exome, or *SMARCA4* Sanger sequencing of 24 predominantly nonfamilial SCCOHT patient tumors and/or germlines collected at multiple institutes [16, 17]. Collections included a research study with a web-based enrollment process at the Translational Genomics Research Institute (TGen) in Phoenix, Arizona, as well as IRB-approved protocols at the Ovarian Cancer Research Program (OvCaRe) tissue bank in Vancouver, British Columbia, and the University of Toronto in Toronto, Ontario, Canada; the Children's Oncology Group at Nationwide Children's Hospital in Columbus, Ohio, USA; and the Hospital de la Santa Creu i Sant Pau at the Autonomous University of Barcelona in Barcelona, Spain.

2.2 Genetic Epidemiology

Familial transmission of cancer can be a particularly powerful setting in which to dissect the genetic basis of a tumor, particularly in the absence of large nonfamilial patient cohorts. Aggregation and segregation studies can elucidate genetic causes and inheritance

patterns that can be further dissected through linkage and association studies. While many common cancers have a complex genetic and environmental etiology (e.g., multiple causative factors in breast cancer including *BRCA* mutations, parity and hormone exposure, and obesity), rare cancers are often driven by more uniform causes (e.g., asbestos in mesothelioma or *RBI* mutations in retinoblastoma). In the case of rare cancers showing patterns of autosomal dominant transmission, a known characteristic of SCCOHT for over two decades, application of NGS using only a few families may be sufficiently powered to detect candidate cancer-driving mutations that may also be common in nonfamilial cases [11, 24–28, 29]. In SCCOHT, 9% (29/312) of total known cases arose from 14 families, clearly indicating a heritable component that, in combination with SCCOHT family pedigrees showing autosomal dominant transmission, was consistent with simple Mendelian patterns of inheritance [10, 11, 18, 24–28, 29]. Finally, cancers with an early onset suggest a genetic basis due to the presence of predisposing germline mutations in a tumor suppressor gene. In contrast to most ovarian cancers, SCCOHT occurs at a median age of 24, often arising in women and girls, with the youngest case diagnosed at 14 months old. Thus, the autosomal dominant transmission of SCCOHT in families strongly supported the hypothesis that this was a monogenic disorder whose basis could be elucidated through NGS in familial and nonfamilial cases.

2.3 Next Generation Sequencing

Next generation sequencing technology has transformed cancer genomics by enabling rapid, cost-effective, and comprehensive evaluation of germline and somatic mutations in large patient cohorts. However, the ability to identify driver mutations depends both on technical features of genomics platforms (e.g., breadth and depth of sequencing), and also on inherent features of the cancer subtype under study (e.g., subtype heterogeneity or heritable fraction). Despite the rarity of the cancer, the existence of affected families along with the assembly of patient cohorts through multi-institutional consortia and the monogenic nature of the disease all contributed to our ability to rapidly identify *SMARCA4* as the central tumor suppressor in SCCOHT. It is also important to note that each of these studies was enabled by sequencing approaches using archival formalin-fixed, paraffin-embedded tissue (FFPE), without which, these cohorts could not have been assembled due to a dearth of fresh frozen tissue. This section outlines the approach and platforms used by the four teams that discovered the genetic basis of SCCOHT:

1. Kupryjańczyk et al. were the first to describe *SMARCA4* mutations in SCCOHT [14]. This group examined *SMARCA4* because histopathology suggested similarity to rhabdoid tumors (RTs), another rare pediatric cancer subtype. Given

that RTs are characterized by mutation and inactivation of the SWI/SNF complex member *SMARCB1*, with reports of occasional *SMARCA4* mutation and loss [30, 31], Kupryjańczyk et al. assessed two SCCOHT tumors for both *SMARCB1* and *SMARCA4* by IHC. They found that both tumors retained *SMARCB1*, but lost expression of *SMARCA4*. They next performed Sanger sequencing of the coding regions of *SMARCA4* on DNA from FFPE and confirmed *SMARCA4* mutations in both cases. The first tumor carried a nonsense mutation resulting in a premature stop codon in addition to loss of heterozygosity, while the second tumor bore both frameshift and nonsense mutations. Based on the anatomical site of SCCOHT, the connection to RTs would have been surprising if not considering their similar pathology. Yet this observation ultimately allowed this group to use a targeted sequencing approach to identify a driver mutation. However, in the absence of familial studies, comprehensive sequencing, or functional biology, it was not clear from this report whether loss of *SMARCA4* was the principle driver of SCCOHT.

2. Witkowski et al. performed whole exome sequencing (WES) using DNA from blood or FFPE tumors of SCCOHT families, with additional sequencing in nonfamilial cases and cases with an unknown family history [18]. WES is an attractive gene discovery platform that balances breadth of sequencing with a focus on functionally annotated genomic regions that are most likely to be clinically relevant. Based on observations of autosomal dominant transmission of SCCOHT, they first sequenced tumor and normal DNA from six affected individuals from three families. In the initial WES of these SCCOHT families, they identified genes containing variants in at least two of the three affected families, narrowing the list to 19 candidates. This list was based on standard filtering that included selection of nonsynonymous and splicing mutations, coding insertion-deletions (indels), and untranslated region mutations, while excluding common variants found in the general population (using 1000 Genomes Project data). Using this approach, *SMARCA4* was the only candidate mutated in all three families. Importantly, sequencing of tumors with matched germlines in the affected cases not only provided increased sensitivity and specificity for calling somatic mutations, but also enabled observation of tumor suppressor behavior. In all three families, germline *SMARCA4* mutations were shared in the germlines of mother and daughter and coupled with a second somatic hit in each tumor—either loss of heterozygosity or mutation of the second allele. This team next used Sanger sequencing to evaluate *SMARCA4* in one additional familial set, as well as 23 individual cases, confirming the presence of

germline mutations in the familial cases and six additional cases, possibly due to de novo germline mutations, undocumented family history, or incomplete penetrance. Finally, loss of SMARCA4 protein expression was confirmed by IHC analysis in SCCOHT tumors, further supporting the role of *SMARCA4* as a tumor suppressor gene in these cancers.

3. Jelinic et al. performed matched tumor/germline sequencing in 12 SCCOHT cases with undocumented family history using a targeted panel of 279 cancer genes [15]. The custom hybrid-capture-based IMPACT (Integrated Mutation Profiling of Actionable Cancer Targets) panel developed at Memorial Sloan Kettering Cancer Center [32] comprises an approach focused more narrowly on the coding regions and introns of genes with a clear link to cancer and, in many cases, with putative links to targeted therapy. Biallelic *SMARCA4* mutations were present in all 12 tumors, and included nonsense, frameshift, and splice-site mutations. Jelinic et al. [15] found only four additional nonrecurrent somatic mutations in cancer genes in these samples, highlighting the low coding mutation burden in SCCOHT and the role of *SMARCA4* as the sole driver in this cancer. They further highlighted the minute probability of identifying by chance the universal mutation of a single gene on this panel through comparative analysis with 4,784 nonhypermuted tumors from The Cancer Genome Atlas dataset, which bore an average of 4.3 somatic mutations in IMPACT panel genes. Finally, *SMARCA4* loss was also confirmed at the protein level in most cases by Western blotting and IHC.
4. Ramos et al. utilized WES, whole genome sequencing, and Sanger sequencing using one SCCOHT cell line and fresh frozen tissue from 12 patients, including nine SCCOHT tumors, and seven germline samples [16]. Based on standard filtering such as that used by Witkowski et al. [18], as well as subtraction of germline variants from matched tumor/germline samples, *SMARCA4* was the only recurrently mutated gene in this dataset, with mutations occurring in six out of nine tumors. Again, these mutations appeared to be largely biallelic and inactivating, and included nonsense, frameshift, and splice-site alterations. IHC was then performed to confirm loss of *SMARCA4* in the corresponding tumors and in an additional nine cases, confirming loss of *SMARCA4* in 82% of SCCOHT tumors. Interestingly, one tumor did not display *SMARCA4* mutations, but instead showed loss of *SMARCB1*. In a follow-up study, Ramos et al. reported on an additional 12 tumors and second cell line, of which, all but one bore mutations in *SMARCA4* [17].

In a recent collaborative meta-analysis incorporating data from all four studies in addition to several new cases, a total of 96 unique, predominantly inactivating (coinciding with loss of SMARCA4 protein by IHC), *SMARCA4* mutations were catalogued from 89 patients [10]. Twenty-six patients bore germline mutations (29%), 34 had only somatic mutations (38%), and germline status was unknown in the remaining 29 patients. Notably, most of the germline carriers had no reported family history and were diagnosed at a slightly younger median age of 21.5 years versus 25.5 years for noncarriers.

2.4 Diagnostic Pathology

While disease gene discovery efforts are often stimulated by clinical observations, research discoveries can also directly impact diagnostic and therapeutic practice. Prior to identification of SMARCA4 in the development of SCCOHT, the differential diagnosis of the cancer was based on age of onset, co-occurrence of hypercalcemia, and relatively nonspecific pathology observations in the absence of definitive immunohistochemical markers. SCCOHT is characterized by nests and cords of diffuse proliferation of small undifferentiated cells containing occasional pseudofollicles and, in about one third of cases, rhabdoid-like cells with abundant eosinophilic cytoplasm [33]. This histologic similarity to RTs led to the first sequencing of *SMARCA4* and discovery of its loss in these tumors [14]. Subsequent sequencing studies have now confirmed that SMARCA4 protein loss accompanies mutation in virtually all SCCOHT cases. Further, we have now shown through IHC analyses that SCCOHTs also universally lack expression of the alternative SWI/SNF ATPase, SMARCA2, which is epigenetically silenced in these tumors [20]. We also assessed over 3,000 primary gynecological tumors for SMARCA4 and SMARCA2, and found that SMARCA4 loss is relatively unique to SCCOHT among gynecological tumors, whereas dual loss of both SWI/SNF ATPases is completely specific for SCCOHT, findings also confirmed by others [19]. Following the accumulation of evidence in favor of *SMARCA4* mutation as an SCCOHT driver, and the concomitant loss of protein in these tumors by IHC, dual loss of SMARCA4 and SMARCA2 are now definitively diagnostic for SCCOHT and have immediate relevance for prognosis and treatment considerations.

2.5 Model Systems

Validation of putative cancer-driving mutations requires use of model systems to establish transformational causality and characterize downstream biology. Furthermore, preclinical models are critical for discovery of new effective treatments. However, model systems for rare cancers are themselves just as rare. Although more than 1,000 immortalized cancer cell lines exist, few of these cell lines are derived from rare cancer subtypes. For example, only two published SCCOHT cell lines exist and, while both lines bear *SMARCA4* mutations and SMARCA2 protein loss, they differ

significantly in morphology, growth characteristics, and drug response [34–36]. Meanwhile, establishment of new cell lines can be challenging with a high failure rate for particular cancer types that is compounded by the scarcity of sufficient fresh tumor material from rare cancers. Many variables must be considered in generating new cell lines, and include method of tissue preparation or cell isolation, cell culture media used, and validation against expected disease phenotypes (see “Guidelines for the use of cell lines in biomedical research” for a complete summary [37]). Phenotypic validation includes examination of known tumor markers, morphological characteristics, and histology and growth characteristics when tumors are xenografted into mice. Finally, in the case of cancers for which the cell of origin remains unknown such as SCCOHT, lines are limited to those derived directly from patient tumors, rather than immortalized or engineered lines derived from normal precursor cells. In lieu of subtype-specific models, alternatives may be used to mimic the genetic context of the disease either through engineering oncogenic mutations (e.g., utilizing recent advances in CRISPR/Cas9 genome editing to create precise genetic alterations in isogenic pairs) into simplified models such as HEK293 cells or immortalized fibroblasts. These models are not ideal, however, as many genetic alterations bear cell context-specific functions. Other options include examining cell line databases for lines bearing mutations similar to the candidate cancer-driving mutations identified in the cancer of interest. The Cancer Cell Line Encyclopedia (Broad Institute), for example, contains genomic (multi-platform DNA and RNA) and pharmacological profiling data on ~1,000 cancer cell lines, as well as visualization and analysis tools for identifying cell lines with similar genetic and drug response profiles [38]. In the case of SCCOHT, both SCCOHT cell lines and substitutes (such as SMARCA4/SMARCA2-null non-small cell lung cancer cell lines) have been utilized for functional and therapeutic studies [15, 20, 35, 36].

Mouse models allow complex modeling of cancer physiology, disease course, and treatment response in living systems with pharmacokinetic and pharmacodynamic constraints. For rare cancers, however, such models may not exist or histologies and genotypes may not be recognized in existing models. Mouse models can be derived by xenografting established cell lines, establishing patient-derived xenografts (PDXs) directly from fresh surgical tissue, or utilizing reverse genetics in syngeneic systems. Xenograft models (whether cell line xenografts or PDXs) can be very useful for study of rare diseases given their relative flexibility and scalability in addition to easily measured and rapidly achieved study endpoints. However, grafted tumors typically lack genetic and cellular heterogeneity, as well as interactions with vasculature and normal stroma that are present in naturally occurring tumors, even when implanted orthotopically. PDXs may better recapitulate tumor

heterogeneity, host interactions, and drug response compared to cell line xenografts, but they also tend to be more difficult to establish and maintain. Finally, syngeneic mouse models can be created through reverse genetics approaches when a candidate gene is known. For many candidate cancer genes, prior knowledge exists about the viability and phenotype of the associated transgenic mouse. If the alteration is embryonic lethal, conditional expression of the mutant gene may be performed in a tissue-specific manner through use of temporally controlled Cre-mediated excision or drug-dependent promoter expression (e.g., tetracycline or tamoxifen inducible systems). In the case of SCCOHT, transgenic mouse models have thus far been of limited utility because homozygous *Smarca4* knockouts are embryonic lethal. Heterozygous *Smarca4* knockouts are also of limited utility in understanding SCCOHT because, although 10% of heterozygous knockouts develop tumors within 1 year, these tumors are mammary carcinomas resulting from haploinsufficiency and genomic instability that bear little resemblance to SCCOHT [39, 40]. Nonetheless, two cell line xenograft and three PDX models have been described [17, 19, 34] that demonstrate similar pathology and growth characteristics as SCCOHT. These models currently power ongoing functional and treatment studies in SCCOHT.

3 Conclusions

As shown by the studies describing the discovery of SMARCA4 mutations in the development of SCCOHT, charting the genetic underpinnings of rare cancers in the post-genomic era is best served by multi-institutional research consortia that employ integration of clinical and research partners, genetic epidemiology, NGS, pathology, and validation in cell and animal models. By determining the genetic basis of rare cancers, we can better diagnose and improve outcomes for these patients, for whom treatment outcomes lag behind those of common cancers.

Acknowledgments

The authors thank the SCCOHT patients, families, clinicians, and support and advocacy groups for their critically important contributions to the IRB-approved study performed at TGen. Our work results from a multi-institutional effort and we thank the many scientists and clinicians who have contributed to this work, as well as Drs. Jeffrey Trent, David Huntsman, and Bernard Weissman for their leadership in these collaborations. Our work has been supported by grants from the National Institutes of Health (R01 CA195670-01), the Anne Rita Monahan Foundation, the Marsha

Rivkin Center for Ovarian Cancer Research, the Ovarian Cancer Alliance of Arizona, the Small Cell Ovarian Cancer Foundation, and philanthropic support to the TGen Foundation.

References

- Greenlee RT, Goodman MT, Lynch CF et al (2010) The occurrence of rare cancers in US adults, 1995-2004. *Public Health Rep* 125:28-43
- Gatta G, van der Zwan JM, Casali PG et al (2011) Rare cancers are not so rare: the rare cancer burden in Europe. *Eur J Cancer* 47 (17):2493-2511. <https://doi.org/10.1016/j.ejca.2011.08.008>
- Bogaerts J, Sydes MR, Keat N et al (2015) Clinical trial designs for rare diseases: studies developed and discussed by the International Rare Cancers Initiative. *Eur J Cancer* 51 (3):271-281. <https://doi.org/10.1016/j.ejca.2014.10.027>
- Boyd N, Dancey JE, Gilks CB et al (2016) Rare cancers: a sea of opportunity. *Lancet Oncol* 17 (2):e52-e61
- Knudson AG (1971) Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* 68(4):820-823
- Friend SH, Bernards R, Rogelj S et al (1986) A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* 323(6089):643-646
- Broadbent E, Topham A, Singh AD (2009) Incidence of retinoblastoma in the USA: 1975-2004. *Br J Ophthalmol* 93(1):21-23
- MacCarthy A, Birch J, Draper G et al (2009) Retinoblastoma in Great Britain 1963-2002. *Br J Ophthalmol* 93(1):33-37
- Scully RE (1979) Tumors of the ovary and maldeveloped gonads. *Atlas of tumor pathology*, vol 2. Armed Forces Institute of Pathology, Washington, DC. ser, fasc 16
- Witkowski L, Goudie C, Ramos P et al (2016) The influence of clinical and genetic factors on patient outcome in small cell carcinoma of the ovary, hypercalcemic type. *Gynecol Oncol* 141 (3):454-460. <https://doi.org/10.1016/j.ygyno.2016.03.013>
- Martinez-Borges AR, Petty JK, Hurt G et al (2009) Familial small cell carcinoma of the ovary. *Pediatr Blood Cancer* 53 (7):1334-1336. <https://doi.org/10.1002/pbc.22184>
- Stephens B, Anthony SP, Han H et al (2012) Molecular characterization of a patient's small cell carcinoma of the ovary of the hypercalcemic type. *J Cancer* 3:58
- Gamwell LF, Gambaro K, Merziotis M et al (2013) Small cell ovarian carcinoma: genomic stability and responsiveness to therapeutics. *Orphanet J Rare Dis* 8:33
- Kupryjanczyk J, Dansonka-Mieszkowska A, Moes-Sosnowska J et al (2013) Ovarian small cell carcinoma of hypercalcemic type - evidence of germline origin and SMARCA4 gene inactivation. A pilot study. *Pol J Pathol* 64 (4):238-246
- Jelinic P, Mueller JJ, Olvera N et al (2014) Recurrent SMARCA4 mutations in small cell carcinoma of the ovary. *Nat Genet* 46 (5):424-426. <https://doi.org/10.1038/ng.2922>
- Ramos P, Karnezis AN, Craig DW et al (2014) Small cell carcinoma of the ovary, hypercalcemic type, displays frequent inactivating germline and somatic mutations in SMARCA4. *Nat Genet* 46(5):427-429. <https://doi.org/10.1038/ng.2928>
- Ramos P, Karnezis AN, Hendricks WP et al (2014) Loss of the tumor suppressor SMARCA4 in small cell carcinoma of the ovary, hypercalcemic type (SCCOHT). *Rare Dis* 2(1):e967148
- Witkowski L, Carrot-Zhang J, Albrecht S et al (2014) Germline and somatic SMARCA4 mutations characterize small cell carcinoma of the ovary, hypercalcemic type. *Nat Genet* 46 (5):438-443. <https://doi.org/10.1038/ng.2931>
- Jelinic P, Schlappe BA, Conlon N et al (2015) Concomitant loss of SMARCA2 and SMARCA4 expression in small cell carcinoma of the ovary, hypercalcemic type. *Mod Pathol* 29(1):60-66
- Karnezis AN, Wang Y, Ramos P et al (2016) Dual loss of the SWI/SNF complex ATPases SMARCA4/BRG1 and SMARCA2/BRM is highly sensitive and specific for small cell carcinoma of the ovary, hypercalcemic type. *J Pathol* 238(3):389-400
- Shah SP, Köbel M, Senz J et al (2009) Mutation of FOXL2 in granulosa-cell tumors of the ovary. *N Engl J Med* 360(26):2719-2729
- Tiacci E, Trifonov V, Schiavoni G et al (2011) BRAF mutations in hairy-cell leukemia. *N Engl J Med* 364(24):2305-2315

23. Choy E, MacConaill LE, Cote GM et al (2014) Genotyping cancer-associated genes in chordoma identifies mutations in oncogenes and areas of chromosomal loss involving CDKN2A, PTEN, and SMARCB1. *PLoS One* 9(7):e101283
24. Ulbright TM, Roth LM, Stehman FB et al (1987) Poorly differentiated (small cell) carcinoma of the ovary in young women: evidence supporting a germ cell origin. *Hum Pathol* 18(2):175–184
25. Peccatori F, Bonazzi C, Lucchini V et al (1993) Primary ovarian small cell carcinoma: four more cases. *Gynecol Oncol* 49(1):95–99. <https://doi.org/10.1006/gyno.1993.1093>
26. Lamovec J, Bracko M, Cerar O (1995) Familial occurrence of small-cell carcinoma of the ovary. *Arch Pathol Lab Med* 119(6):523–527
27. Longy M, Toulouse C, Mage P et al (1996) Familial cluster of ovarian small cell carcinoma: a new mendelian entity? *J Med Genet* 33(4):333–335
28. Distelmaier F, Calaminus G, Harms D et al (2006) Ovarian small cell carcinoma of the hypercalcemic type in children and adolescents: a prognostically unfavorable but curable disease. *Cancer* 107(9):2298–2306. <https://doi.org/10.1002/cncr.22213>
29. McDonald JM, Karabakhtsian RG, Pierce HH et al (2012) Small cell carcinoma of the ovary of hypercalcemic type: a case report. *J Pediatr Surg* 47(3):588–592. <https://doi.org/10.1016/j.jpedsurg.2011.12.004>
30. Eaton KW, Tooke LS, Wainwright LM et al (2011) Spectrum of SMARCB1/INI1 mutations in familial and sporadic rhabdoid tumors. *Pediatr Blood Cancer* 56(1):7–15
31. Hasselblatt M, Gesk S, Oyen F et al (2011) Nonsense mutation and inactivation of SMARCA4 (BRG1) in an atypical teratoid/rhabdoid tumor showing retained SMARCB1 (INI1) expression. *Am J Surg Pathol* 35(6):933–935
32. Won HH, Scott SN, Brannon AR et al (2013) Detecting somatic genetic alterations in tumor specimens by exon capture and massively parallel sequencing. *J Vis Exp* (80):e50710
33. Richard Dickersin G, Kline IW, Scully RE (1982) Small cell carcinoma of the ovary with hypercalcemia: a report of eleven cases. *Cancer* 49(1):188–197
34. Otte A, Gohring G, Steinemann D et al (2012) A tumor-derived population (SCCOHT-1) as cellular model for a small cell ovarian carcinoma of the hypercalcemic type. *Int J Oncol* 41(2):765–775. <https://doi.org/10.3892/ijo.2012.1468>
35. Otte A, Rauprich F, Hillemanns P et al (2014) In vitro and in vivo therapeutic approach for a small cell carcinoma of the ovary hypercalcemic type using a SCCOHT-1 cellular model. *Orphanet J Rare Dis* 9:126. <https://doi.org/10.1186/s13023-014-0126-4>
36. Otte A, Yang Y, von der Ohe J et al (2016) SCCOHT tumors acquire chemoresistance and protection by interacting mesenchymal stroma/stem cells within the tumor microenvironment. *Int J Oncol* 49(6):2453–2463
37. Geraghty RJ, Capes-Davis A, Davis JM et al (2014) Guidelines for the use of cell lines in biomedical research. *Br J Cancer* 111(6):1021–1046. <https://doi.org/10.1038/bjc.2014.166>
38. Barretina J, Caponigro G, Stransky N et al (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483(7391):603–607
39. Bultman S, Gebuhr T, Yee D et al (2000) A *Brg1* Null Mutation in the Mouse Reveals Functional Differences among Mammalian SWI/SNF Complexes. *Mol Cell* 6(6):1287–1295
40. Bultman S, Herschkowitz J, Godfrey V et al (2007) Characterization of mammary tumors from *Brg1* heterozygous mice. *Oncogene* 27(4):460–468

Chapter 21

The Rise and Fall and Rise of Linkage Analysis as a Technique for Finding and Characterizing Inherited Influences on Disease Expression

Ettie M. Lipner and David A. Greenberg

Abstract

For many years, family-based studies using linkage analysis represented the primary approach for identifying disease genes. This strategy is responsible for the identification of the greatest number of genes proven to cause human disease. However, technical advancements in next generation sequencing and high throughput genotyping, coupled with the apparent simplicity of association testing, led to the rejection of family-based studies and of linkage analysis. At present, genetic association methods, using case-control comparisons, have become the exclusive approach for detecting disease-related genes, particularly those underlying common, complex diseases. In this chapter, we present a historical overview of linkage analysis, including a description of how the approach works, as well as its strengths and weaknesses. We discuss how the transition from family-based studies to population comparison association studies led to a critical loss of information with respect to genetic etiology and inheritance, and we present historical and contemporary examples of linkage analysis “success stories” in identifying genes contributing to the development of human disease. Currently, linkage analysis is re-emerging as a useful approach for identifying disease genes, determining genetic parameters, and resolving genetic heterogeneity. We posit that the combination of linkage analysis, association testing, and high throughput sequencing provides a powerful approach for identifying disease-causing genes.

Key words Linkage analysis, Association analysis, GWAS, Recombination fraction, Genetic heterogeneity, Common disease, Phenotype, BRCA1, Crohn’s disease, Epilepsy

1 An Historical Overview of Linkage Analysis

At one time, linkage analysis was the primary statistical genetic approach for identifying loci underlying Mendelian and complex diseases. Despite dramatic improvements in next generation sequencing and high throughput genotyping techniques, linkage analysis has successfully identified the greatest number of genes proven to cause human disease compared to any other method.

With the completion of the human genome project, techniques such as genome-wide association studies (GWAS), which identify

associations between markers (typically, single nucleotide polymorphisms or SNPs) and disease, quickly eclipsed linkage analysis in popularity. There are several reasons for the widespread embrace of GWAS, including the ease of data collection (cases and controls, rather than families), simplicity of analytic methods (e.g., chi-squared 2×2 tables), and the belief that alleles contributing to disease must be detectable given large enough sample sizes and enough markers. Subsequently, whole exome (WES) and whole genome sequencing (WGS) became feasible approaches, and are becoming increasingly popular, because of the (theoretical) power to identify disease-related mutations.

Many investigators believed that comparing allele frequency differences between cases and controls using GWAS or WES/WGS would lead to the identification of important disease genes. However, problems inherent in studying common, complex diseases have left many findings of genetic association unreplicated. These problems include heterogeneity, misdiagnosis, and misclassification of phenotypes. While GWAS have yielded highly significant p -values with very large sample sizes, the corresponding odds ratios have been mostly <1.5 . Thus, when GWAS successfully identified disease-associated alleles, the actual impact of the variants on disease risk was very low. These loci are defined as “susceptibility genes,” or genes that increase risk, as opposed to genes that, when mutated, actually cause disease [1].

In contrast, linkage analysis has a record of accurately and definitively identifying loci with major effects on disease etiology. Linkage assesses co-segregation of disease with marker alleles within families and yields information not only about disease gene location, but also about disease inheritance, penetrance, heterogeneity, gene–gene interactions, movement of genomic regions across generations, and the manner in which those regions are related to the presence of disease. All of these attributes of linkage lie beyond the reach of population-based association analysis. Contrary to popular belief, linkage analysis is indeed useful for identifying genes underlying common diseases, although the efficiency and apparent transparency of the (perhaps, simplistic) statistical tests used for genetic association analysis quickly made association analysis the preferred approach. Consequently, the ability to derive information about the genetic inheritance of a disease was sacrificed [2].

2 How Linkage Analysis Works

Linkage analysis identifies genomic regions that contain disease-predisposing genes by quantifying the cosegregation of a phenotype (i.e., a disease) with alleles at a marker locus. Linkage determines whether marker alleles within a family are inherited together with a disease through generations. Linkage relies on the

phenomenon of recombination, that is, if loci are close together on the same chromosome, alleles at those neighboring loci will segregate together in families more often than expected by chance. The farther apart two loci are on the same chromosome, the more likely that a recombination event during meiosis will cause the alleles at these loci to end up on different chromosomal strands, and will not be found together in subsequent generations. Likewise, alleles at loci located on different chromosomes segregate independently of one another.

2.1 LOD Scores

Linkage analysis compares the likelihood that an allele at a marker locus segregates with the disease (or phenotype) with the likelihood that alleles at the marker locus and the disease segregate independently. The statistical measure for determining if two loci are linked is reported as a logarithm of the odds for linkage (or LOD score), which is a function of the recombination fraction. The accepted cut-off value for statistically significant evidence of linkage is a LOD score over 3 [3], which is interpreted as the data showing 1000 times more support for the existence of linkage than for non-linkage (that is, marker and disease segregate independently). LOD scores < -2 are evidence against linkage, implying that non-linkage is 100 times more likely than linkage.

2.2 Phenotyping and Heterogeneity

The success and utility of *any* genetic study are dependent on precise phenotyping. For example, the majority of common and complex diseases comprise many clinically similar disease subtypes that are caused by different etiological factors, both genetic and non-genetic. Diabetes is a prime example of the complexities involved in phenotyping. Aside from distinguishing type 1 diabetes (T1D) from type 2 diabetes (T2D), etiologically separate forms of diabetes can be further distinguished by even narrower phenotype definitions, such as latent autoimmune diabetes in adults (LADA), maturity onset diabetes of the young (MODY), gestational diabetes, neonatal diabetes mellitus (NDM), and mitochondrial diabetes mellitus [4]. Similarly, in epilepsy, the phenotypic syndromes that fall under the classification of idiopathic generalized epilepsy (IGE) (now often referred to as Genetic Generalized Epilepsy) include childhood-onset absence, juvenile-onset absence, juvenile myoclonic epilepsy (JME), epilepsy with generalized tonic-clonic seizures, and awakening grand mal [5]. Before the genetic contributions to disease pathogenesis can be understood, diagnoses must be based on clinical presentation, rather than on the genetic etiologies of these conditions. However, to make genetic studies meaningful, the diagnoses must be as narrowly drawn as possible. Including too many etiologically different forms of disease in the data leads to a state of heterogeneity, in which multiple loci independently cause similar phenotypes that are lumped together as one disease. As a result, any evidence of the effect of any one of those

genes would be obscured. By narrowing the phenotype definition, it may be possible to gain insight into the genetic etiologies and correctly redefine some of these complex diseases, as the examples below will show. Most GWAS attempt to overcome this problem with extremely large study samples, and, in fact, it may be possible to detect statistically significant evidence for association with a large enough data set. However, without the detailed clinical information that would allow one to follow up on carriers of the risk allele, compared with affected non-carriers, further analysis of the GWAS findings are problematic.

Unlike association analysis, linkage analysis can identify the existence of genetic heterogeneity and be used to classify who is affected with which form of disease.

2.3 Possible Confounders in Linkage Analysis

2.3.1 Heterogeneity

Genetic heterogeneity is problematic for *any* genetic study, but there are two types of genetic heterogeneity that influence results of genetic studies.

The first type is allelic heterogeneity, in which different alleles at a locus cause the same or similar disease (e.g., Duchenne's and Becker muscular dystrophy). Linkage analysis is robust to allelic heterogeneity because the identity of the marker alleles is irrelevant; the only requirement is that the marker alleles are inherited together with the disease in a family. Put another way, linkage is based on identifying, not alleles, but *loci*. In contrast, allelic heterogeneity is a major confounder in association analysis, which tests whether specific *alleles* are more frequent in one population than in another. Many different alleles at the marker locus may be in the population. If so, even if several different marker alleles are "close" to (i.e., in linkage disequilibrium with [see below]) the disease-causing variant, it may not be possible to detect population allele frequency differences with any one of them [6]. There is no technique that can compensate for this problem in association analysis.

The second type of heterogeneity is locus heterogeneity, which arises when disease phenotypes result from different genetic etiologies, yet are clinically indistinguishable. Because linkage assesses co-segregation of marker alleles and disease, a disease locus (i.e., locus A) for which there has been no recombination between the marker loci and the disease locus (within families) will show co-segregation of the disease with alleles near the marker locus. If family structures are introduced into the data set where one locus exists (i.e., locus B) that produces a disease that is phenotypically identical to another produced by locus A, but with a different genetic etiology, the result is the creation of recombinations between the "disease" and the markers surrounding locus A, because locus A is not causing the disease in those heterogeneous families. The recombinations then "dilute" the evidence for linkage at locus A because simple LOD scores assume homogeneity in the data set. Locus heterogeneity is also a problem for association analysis for reasons similar to those for linkage. Heterogeneity

dilutes the signal at the true disease locus by introducing alleles not in linkage disequilibrium (LD) with the disease locus, as some subjects have a disease caused by a genetically distant locus, not associated with the allele being tested.

While association analysis has no way to account for locus heterogeneity, linkage methods include a heterogeneity LOD score (i.e., *hetlod* or *HLOD*). The *HLOD* calculates the likelihood of a mixture of two distributions of LOD scores, one centered on a positive LOD score value and another on a negative one. In calculating the *HLOD*, one estimates the value of α (the proportion of linked families), while also estimating the recombination fraction (θ) [7]. Thus, loci cannot only be detected in the presence of heterogeneity, but the percentage of families in the data set that are linked compared to those that are not linked can be estimated, and these linked families can be identified.

2.3.2 Penetrance

Penetrance is the probability of manifesting the trait (or disease) phenotype, given the presence of the trait genotype. For complex traits, penetrance is usually incomplete, meaning that not everyone who has the disease-causing variant manifests the disease. Penetrance can depend on many factors, such as age and sex. For example, Huntington's disease is an inherited autosomal dominant disorder caused by the expansion of a CAG trinucleotide within the *HTT* gene on chromosome 4. At birth, there is zero penetrance of the disease, but, in later life, complete penetrance (i.e., everyone who has the CAG repeat has the disease) develops [8]. Reduced penetrance lowers the informativeness of families for linkage, but it will not undermine the ability to detect linkage; it will only increase the size of the data set needed. This is different from the effect of locus heterogeneity, which can hide a linkage signal and may not be resolved simply by increasing the sample size.

Reduced penetrance can be caused by environmental (e.g., exposure to a chemical or allergen) or genetic factors. Penetrance modified by genetic factors can occur when two loci are required for disease expression, that is, when at least one other disease locus *interacts* with the first locus, such that the probability of disease expression is dependent on the presence of disease alleles at both loci. Individuals carrying the disease allele at only one locus will appear to be nonpenetrant. However, linkage can identify both loci and assess the interaction between them [9, 10]. Linkage analysis can also be used to estimate the penetrance, whatever the cause [11, 12]. Genetic association analysis cannot control for penetrance levels.

2.3.3 Mode of Inheritance

To perform linkage analysis, a mode of inheritance (MOI), or the structure of how the disease is inherited, must be specified. When the MOI is not known, or if it is misspecified, the power to detect linkage can be reduced. To get around this problem, nonparametric

linkage techniques, such as the affected sibling-pair (ASP) and affected-only methods, were developed. These methods do not require specifying a mode of inheritance. Many researchers prefer to use these nonparametric methods when the MOI is not known, even though LOD-score analysis has greater power to detect linkage [13]. However, parametric linkage analysis can be used to obtain a useful estimate of the MOI at the detected locus, determining whether the locus is dominantly, recessively, or additively inherited. It can also estimate the penetrance at the locus, whether due to environmental or genetic factors. With nonparametric linkage analysis, however, an HLOD cannot be calculated and penetrance cannot be estimated. The belief is still widespread that nonparametric analyses are superior because no MOI is assumed. However, Whittemore [14] showed that all nonparametric tests correspond to some unknown, and likely unknowable, assumed MOI. As a result, there are advantages to assuming first a dominant, then a recessive mode of inheritance, and choosing the higher of the two values as the best estimate for the mechanism of inheritance [15]. Nonetheless, sophisticated nonparametric methods are both common and reliable. In contrast, mode of inheritance is generally not taken into account in genetic association analyses.

2.3.4 *Sporadic Cases*

Sporadic disease has a clinical presentation similar to a genetic form of disease, but lacks a heritable component. Nongenetic forms of disease can confound linkage results by reducing the signal derived from affected families. While linkage analysis is relatively robust to even notable frequencies of sporadic disease as long as families with more than one affected member are selected [16]. Association analysis, on the other hand, cannot account for confounding by nongenetic cases.

2.3.5 *Gene–Gene Interaction*

In common, complex diseases, detecting gene–gene interaction is not only a challenge, but also is a major priority for human genetics studies. To identify interactions, most approaches start with marker-disease associations identified in GWAS, but many believe that testing for simultaneous multiple associations is an unproductive approach to identifying interaction-based genetic effects [9]. These association-based interaction tests are extremely weak because they rely only on allele frequency differences, use data only at the population level, and do not utilize inheritance information available from family data. As a result, finding gene–gene interactions with GWAS data is problematic, especially where so many other confounders can obscure a single gene association or linkage signal. There are several methods aiming to identify statistical interactions between loci from genetic association studies [17–19], but there are disadvantages to these approaches. If allelic heterogeneity is present, the power to detect association when using the single-

locus testing approach is dramatically reduced. Testing for allelic interactions at multiple loci is even more difficult and reduces power further. Interaction testing requires enormous sample sizes due to the required multiple testing correction. Lastly, in the case of statistically significant evidence supporting an interaction between two alleles that show non-significant associations with disease on their own, the biological significance of this scenario is unclear [9]. In contrast, there are some novel approaches that can be used to detect gene–gene interaction that take advantage of family data with affected individuals [9, 10] and which have proven successful with real data [20].

3 How Association Analysis Works

The goal of genetic association studies is to detect a difference in allele frequencies between two populations, cases and controls, at a specific locus (or thousands to millions of SNP loci). Statistically significant differences between the frequencies of a marker allele in cases compared to controls suggests that the gene in which the SNP is found, or a nearby gene with variants in linkage disequilibrium with that SNP, affects the expression of the disease [21].

3.1 True Association

When alleles are found together on the same DNA strand *in the population* (as opposed to members of a family), they are said to be in linkage disequilibrium (LD). An association between two loci can be detected because of LD between alleles at two loci. LD can be quantified using various measures. In complete LD, two alleles are always found together in the population, while in the absence of LD, the frequency of the two alleles being found together reflects their population frequencies [22, 23]. Association studies are based on the idea that the marker allele is situated on the same DNA strand as the disease allele and that the two loci (disease and marker) are close enough to each other that few recombination events have occurred between them since the mutations that created those alleles occurred (Fig. 1). With the passage of generations, recombination events can eliminate LD between markers [24]. This means that a specific marker allele may eventually cease to be found on the same DNA strand with the disease allele with greater frequency than any other marker allele. A statistically significant difference in the marker allele frequency found in cases compared with that found in controls provides evidence for association, assuming no methodological issues are influencing the results. A statistically significant association can be due to a direct association, in which the marker alleles are indeed disease causing, or may simply reflect LD between marker and causal alleles.

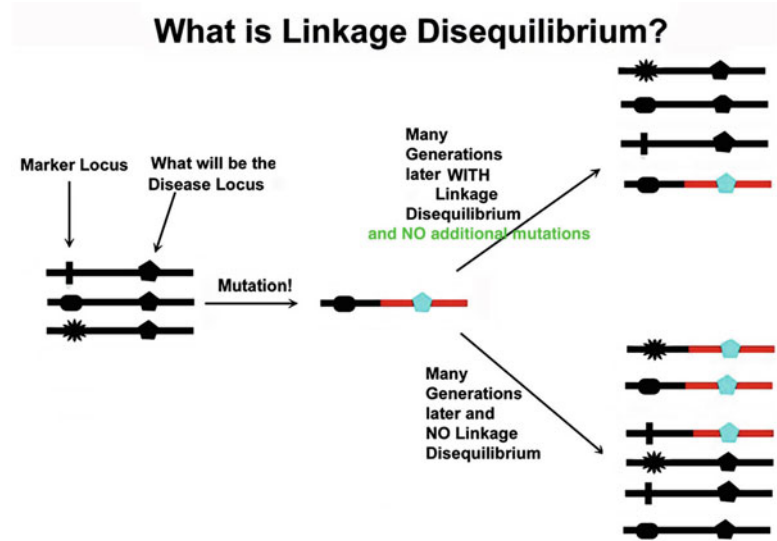


Fig. 1 Principle of linkage disequilibrium (LD). When a new mutation occurs in a gene, the disease allele that results is flanked by a specific combination of marker alleles. Many generations later, in the absence of recombination, the mutation will remain surrounded by the same marker alleles (assuming none have mutated), and all the alleles on this stretch of DNA (i.e., the haplotype) will be in LD. However, in the presence of recombination, the sequence flanking the disease allele will be shuffled so that the mutation is then found with a different combination of markers relative to the disease-related allele

3.2 Association due to Population Stratification

There is another way that one can detect association independent of disease. If cases and controls come from two different populations, that is, if the allele frequencies of the marker alleles are different just by virtue of different evolutionary histories, then associations that have nothing to do with the disease will be detected. This is called “population stratification” [25, 26]. It is a critical issue in association studies, especially for large data sets comprised of individuals recruited from geographically different locations, or when cases and controls derive from different ethnic backgrounds. The only way to reliably test for an association is when cases and controls are drawn from the same underlying population, specifically, one of shared ancestry. Knowledge of these population differences is the basis for Ancestry Informative Markers (AIMs), which are used to assess the origins of an individual or group using population-specific markers. There are techniques for compensating for population stratification, but it remains a possible confounder because there can be subtle differences in populations that may only be detected with very large sample sizes, the kind of sample sizes currently favored in GWAS to ensure highly significant *p*-values.

3.3 Association Analysis Versus Linkage Analysis

One of the strengths of association studies is the ability to detect genes with modest effects. However, if a sample size is large enough, variants can be detected that are neither necessary, nor sufficient, for disease expression [1]. The disease-affecting genotypes found through association studies generally have low odds ratios (OR), and thus, exert only minor effects on the disease prevalence. Weaknesses inherent in association studies have made it difficult to obtain statistically significant positive findings and hinder replication. Population stratification can lead to increased false positives (type I error) and false negatives (type II error). False negative results can arise when the sample size is relatively small, and there is insufficient power to detect association.

While association analysis can detect alleles that increase disease susceptibility, those alleles may not be necessary for disease expression and do not consistently cotransmit with the disease within families. Unless a large degree of the disease risk is conferred by an allele, linkage will be unable to detect such loci. On the other hand, if a gene locus has a major effect on disease expression and segregates with disease in families, then linkage analysis will be able to identify the chromosomal region on which it is located [1]. The disease gene locations identified by linkage analysis can be large, extending over much longer genomic regions compared to the short distances in which LD is in effect. Linkage studies are therefore able to detect disease-related genes in much larger regions than association methods, genomic sections ranging from approximately 2000–20,000 kb. In association studies, frequency differences can be detected in approximately 0–100 kb regions, that is, the most likely range for LD to be in effect [27]. Linkage methods are seldom precise enough to identify a particular gene; rather a linkage finding implicates a wide region that is linked to disease. Thus, association studies in a critical region identified through linkage is generally the next step in identifying a disease-causing gene.

4 Examples of Linkage Success in Common Disease

Although linkage analysis has acquired the unjustified reputation of being unable to detect genes for common, complex diseases, there are actually many examples of its success in identifying disease loci. In this section, we present two common, complex diseases where linkage analysis was successful in identifying disease genes.

4.1 Crohn's Disease

Crohn's disease (CD) is a common, heritable chronic inflammatory disease of the gastrointestinal tract. Many twin studies demonstrated familial clustering of the disease [28–31], and early genetic epidemiology studies showed that inherited factors likely contribute to CD susceptibility [32–35]. The first linkage analysis for CD was performed prior to the sequencing of the human genome, and

utilized widely spaced microsatellite markers [36]. The maximum two-point LOD score ($\text{LOD} = 2.04$) was detected at marker D16S409 on chromosome 16. Additional markers were genotyped in pooled family panels in the 40 cM region flanking D16S409, most of which showed linkage to the locus (named IBD1).

Subsequent linkage studies were performed, but with inconsistent findings [37–42]. In 2001, datasets from 12 centers spanning three continents were combined to achieve greater statistical power [43]. The combined sample size consisted of 613 Caucasian nuclear families, containing two or more offspring with IBD. Family members were diagnosed with CD, ulcerative colitis (UC), or both. Linkage analysis using this study sample identified a peak LOD score of 4.96 near IBD1. However, linkage analysis restricted to families with a CD diagnosis yielded a $\text{LOD} = 5.79$, while no evidence for linkage was observed in this region in families with a UC diagnosis, indicating that IBD1 was, very likely, a locus unique to CD, despite the similar clinical presentations shared between CD and UC.

The strength of these linkage studies is their success in replicating the IBD1 locus and demonstrating the specificity of that gene to CD. The results also illustrate many of the general problems of genetic studies, especially the potential for heterogeneity and imprecise phenotyping that confound results and lead to lack of replication. In the combined analysis, careful phenotyping was crucial in identifying the IBD1 locus. When families with only UC or mixed UC and CD families were analyzed separately, there was no evidence for linkage. While CD and UC are both common, complex genetic disorders affecting the gut, and present with overlapping phenotypes, their pathologies and genetic etiologies are distinct. Linkage analysis was able to genetically distinguish between the two conditions.

4.2 Breast Cancer

Linkage analysis in breast cancer was particularly effective in identifying genes of even modest effect sizes and yielded results that have subsequently had a major impact on public health.

In a landmark study, Hall et al. used linkage analysis to localize a breast cancer gene to chromosome 17q21 [44]. The authors used 23 extended families with 146 cases of breast cancer, corresponding to 329 Caucasian relatives. Patients in these families were characterized by a relatively young age at diagnosis, frequent bilateral disease, and occurrence of disease in males, all of which are characteristic of familial, but not sporadic, breast cancer. A strict phenotyping definition was used in which all histologic types of invasive breast cancer were designated as affected. Multiple approaches to evaluating linkage were used, including single and multipoint LOD score methods, and affected sib-pair nonparametric linkage analysis.

The authors found a maximum two-point LOD score of 3.28 at locus D17S74, but only 40% of families contributed to the evidence for linkage, indicating the presence of heterogeneity. Examination of factors such as age at first pregnancy, number of children, prevalence of fertility problems, use of oral contraceptives, and age at menopause, showed that the only difference between linked and unlinked families was age at breast cancer diagnosis. When seven families with a mean age of diagnosis ≤ 45 years were examined, the LOD score at D17S74 increased to 5.98. In contrast, an analysis including only families with late-onset breast cancer demonstrated evidence against linkage. Thus, heterogeneity was resolved using age of disease onset as a phenotypic criterion, showing that a gene for susceptibility to early-onset breast cancer was located in this chromosomal region.

Findings of linkage for both early-onset breast cancer and ovarian cancer to the same chromosomal region were subsequently confirmed in three large pedigrees [45]. In this study, the maximum two-point LOD score for linkage of breast and ovarian cancer families to locus D17S74 was 2.20. Combined results from this study and the initial one thus strongly implicated the chromosome 17q12-q23 region in early-onset breast cancer and ovarian cancer. These findings were later confirmed in a large-scale study of nearly 300 families ascertained for breast or breast-ovarian cancer from 13 international research groups [46]. This locus was officially designated “BRCA1” [47].

Thus, using family data, these studies demonstrated the localization of a breast cancer susceptibility gene using linkage analysis, work subsequently confirmed using large pedigree data, and further solidified through a large collaborative study. The work also showed that BRCA1 may account for the majority of early-onset breast cancer and ovarian cancer, but the existence of heterogeneity indicated that other genes likely predispose to the disease. Similarly, while BRCA1 accounted for a large proportion of early-onset breast and ovarian cancer, the locus conferred only modest effects for later age-of-onset familial disease. The ability of these studies to replicate and confirm findings is partly due to consistent and strict phenotyping definitions. Also, while evidence for a gene from any genetic study may be diluted due to the presence of genetic heterogeneity, linkage analysis has the ability to parse linked and unlinked families. Again, while allelic heterogeneity does not influence linkage results, both locus and allelic heterogeneity can severely negatively affect genetic association results.

4.3 Lessons from the CD and BRCA1 Stories

The linkage examples described in this chapter were in large part successful because the disease genes involved are necessary for disease expression (though perhaps not sufficient), not merely susceptibility genes. Although the initial sample size in the CD study was small, the IBD1 locus was strongly linked to disease. Genetic

heterogeneity played a crucial role in showing the specificity of IBD1 for CD, but not UC. By observing the proportion of linked and unlinked families, and having comprehensive phenotype information, it was possible to identify those CD families showing linkage to the IBD1 locus. In the breast cancer example, the genetic effect of BRCA1 was strong enough that linkage was detected using even a small sample of families, suggesting that the locus is a major contributor to disease expression, despite the presence of genetic heterogeneity. By using linkage analysis and rich phenotypic information, the authors were able to stratify families based on age of onset to localize the gene underlying breast cancer. When this disease was analyzed using a much larger sample size, linkage was strongest, again, for early age of onset and for breast and ovarian cancer families. For some of the breast cancer-only families, other, unknown loci appear to predispose to the disease.

5 The “Fall” of Linkage and the Rise of Genetic Association Analysis

Despite the success of linkage analysis in identifying disease loci, its popularity has been eclipsed by association methods. There were a number of reasons for the ready adoption of association analysis, chief among them being the idea that common diseases were caused by common variants, a phenomenon known as the “common disease-common variant” hypothesis. Association analyses were expected to be more powerful than linkage approaches based on the assumption that identification of causative alleles through LD would be easy, because one or more of those SNP markers covering the genome would be in LD with disease alleles [48]. However, genotyping $>10^5$ – 10^6 SNP markers over the entire genome exacts a large multiple testing correction to compensate for type I error. Consequently, GWAS requires large sample sizes to detect a statistically significant effect.

SNPs identified by GWAS appear, in general, to act as modest risk alleles, implying that they interact with other genetic or environmental factors to cause disease. Additionally, most of these SNPs are located outside of exons, creating speculation about the functionality of these variants. For example, of the 20 risk loci identified by GWAS for Alzheimer’s disease, none of the most strongly associated SNPs are located in coding regions [49]. GWAS of T2D have identified 76 susceptibility loci, with effect sizes so small, they only explain ~6% of disease risk and 10–20% of disease heritability [50]. Likewise, while GWAS has identified more than 180 loci associated with obesity traits, heritability remains unaccounted for [50].

Association analysis would probably have not been able to identify the IBD1 and BRCA1 loci. Both allelic and locus heterogeneity confound true associations between markers and disease,

but the biggest limitations of association analysis are its inability to use family data and its general lack of consideration of family clinical information. In the BRCA1 story, the observations that all affected family members had early-onset disease and that ovarian cancer was also a familial phenotype were crucial for identifying the locus. If this study had been undertaken using a GWAS approach, neither family data, nor phenotypic details, would have been collected, if for no other reason than the numbers of subjects involved, a number often in the tens of thousands, preclude obtaining detailed clinical information or family histories. Typically, a diagnosis of “breast cancer” or “schizophrenia” or “fever” is considered enough of a phenotype with which to identify genetic etiology in GWAS. There is a belief that electronic medical records (EMR) will make detailed phenotype information available for GWAS; however, our experience is that EMR tends to be less detailed than older written records. The main concern of the medical establishment, after all, is treatment, and without a treatment-oriented impetus, the detailed records needed as a basis for research on large samples of patients are unlikely to become part of normal recordkeeping.

In sum, association analysis can be a powerful tool when used in the correct context. Once a disease locus is identified by linkage, allelic association analysis represents a logical next step for locating the specific gene. The next section illustrates how combining careful phenotyping, family data, linkage analysis, association testing, and sequencing can identify of causal gene for a common disease.

6 Recent Successes of Linkage Analysis

An example of the power of combined linkage and association analysis can be found in the genetic characterization of a common childhood epilepsy syndrome known as rolandic epilepsy (RE). This disease affects approximately 1 in 2500 children, presenting with onset of seizures between 4 and 12 years of age [51, 52]. RE is not only considered one of the most common idiopathic epilepsies of childhood, but is also most impacted by inherited factors [53, 54]. The symptoms of RE overlap with those for severe epilepsy syndromes (i.e., atypical benign partial epilepsy, benign occipital epilepsy, and Landau-Kleffner syndrome), emphasizing the need for careful phenotyping. All patients share the defining electroencephalographic (EEG) abnormality of centrotemporal sharp waves (CTS). CTS is an EEG characteristic that is also observed in children with developmental disorders [55–57]. This suggests that CTS may be a marker for widespread neurodevelopmental abnormalities, rather than a condition specific to RE. Strong clustering of developmental disorders in RE families has been observed [53], and CTS in RE appears to be inherited as an autosomal dominant trait [54].

Genome-wide linkage analysis for RE in carefully diagnosed families identified the highest multipoint LOD score (4.30) at marker D11S914 [58]. No evidence of heterogeneity was observed in the region of linkage, which spanned over 13 cM. Forty-four SNPs across this region were genotyped and three of these, located in the ELP4 gene, were associated with RE. Joint analysis using the discovery sample and a replication panel provided strong evidence that ELP4 variants, rs964112 and rs986527, were associated with CTS in RE families. Subsequent sequencing of the promoter, exons, and flanking regions of the ELP4 gene, found no enrichment of coding polymorphisms, indicating that the causative mutation may exist in noncoding regions. An independent analysis using a larger study sample and a higher resolution genotyping array found evidence for association with rs662702 in the 3' untranslated region of the PAX6 gene [59]. The variant allele was present in 14% of CTS patients and 7.6% of controls, and after adjusting for sex and population structure, homozygosity of this allele conferred 12-fold greater odds of CTS.

Linkage analysis was also used to tie the ELP4-PAX6 gene to a wider phenotype. Verbal dyspraxia [60] has been observed in families of RE patients, even those unaffected with RE or CTS. When these subjects were classified “affected” for the linkage analysis, the evidence for linkage to the RE locus rose from 3.2 to 7.5. It is important to emphasize that, as with the breast cancer findings, being able to define which phenotypes will help differentiate different etiologies and thus, resolve heterogeneity, is one of the most clinically important advantages of linkage analysis because the information directly impacts diagnosis and treatment.

These studies of RE demonstrate a successful paradigm for how gene identification using family studies in common, complex diseases should work. First, a well-described phenotype was necessary, supplemented with detailed family medical histories. Second, a genome-wide linkage analysis identified a genomic region that was significantly linked to a subclinical phenotypic marker (CTS) that was assessed in members of families identified through an RE patient. Subsequently, specific alleles were found to be associated with CTS. Then, by sequencing the linked region, a specific allele in a nonexonic region was implicated in disease pathogenesis. This sequence of steps utilized the strengths of each methodology to narrow in on the gene and allele influencing disease expression.

7 Conclusions

Linkage analysis and family studies are re-emerging as important and useful approaches for identifying genes underlying common, complex diseases. Linkage analysis using family data can yield important information about inheritance and is able to account

for important confounders that can obscure findings in any genetic study, namely, genetic heterogeneity, penetrance, mode of inheritance, sporadic cases, and gene–gene interaction. The ability to account for these confounders creates an important advantage over genetic association analysis, WES, and WGS but, thus far, the genetic research community has failed to take advantage of the enormous wealth of information inherent in family data. Rather, over the past decades, efforts have been focused on collecting population-level data that may have little to do with disease inheritance, and which use analysis approaches that are unable to take into account important confounders inherent in all genetic studies.

Implicit in our title is the belief that the eclipse of family studies and linkage studies by association analyses will eventually come to an end. By coupling linkage analysis with the depth of coverage that genetic sequencing offers, we may be able to identify, with greater certainty and rapidity, genetic mutations that cause complex human diseases.

References

- Greenberg DA (1993) Linkage analysis of “necessary” disease loci versus “susceptibility” loci. *Am J Hum Genet* 52:135–143
- Vieland VJ, Devoto M (2011) Next-generation linkage analysis. *Hum Hered* 72:227
- Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 11:241–247
- Flannick J, Johansson S, Njolstad PR (2016) Common and rare forms of diabetes mellitus: towards a continuum of diabetes subtypes. *Nat Rev Endocrinol* 12:394–406
- Greenberg DA, Stewart WC (2012) How should we be searching for genes for common epilepsy? A critique and a prescription. *Epilepsia* 53(Suppl 4):72–80
- Rodriguez-Murillo L, Greenberg DA (2008) Genetic association analysis: a primer on how it works, its strengths and its weaknesses. *Int J Androl* 31:546–556
- Ott J (1999) *Analysis of human genetic linkage*, 3rd edn. Johns Hopkins University Press, Baltimore, MD
- Mielcarek M (2015) Huntington’s disease is a multi-system disorder. *Rare Dis* 3:e1058464
- Corso B, Greenberg DA (2014) Using linkage analysis to detect gene–gene interaction by stratifying family data on known disease, or disease-associated, alleles. *PLoS One* 9:e93398
- Hodge SE, Hager VR, Greenberg DA (2016) Using Linkage Analysis to Detect Gene–Gene Interactions. 2. Improved Reliability and Extension to More-Complex Models. *PLoS One* 11:e0146240
- Greenberg DA (1989) Inferring mode of inheritance by comparison of lod scores. *Am J Med Genet* 34:480–486
- Greenberg DA (1990) Linkage analysis assuming a single-locus mode of inheritance for traits determined by two loci: inferring mode of inheritance and estimating penetrance. *Genet Epidemiol* 7:467–479
- Greenberg DA, Abreu PC (2001) Determining trait locus position from multipoint analysis: accuracy and power of three different statistics. *Genet Epidemiol* 21:299–314
- Whittmore AS (1996) Genome scanning for linkage: an overview. *Am J Hum Genet* 59:704–716
- Greenberg DA, Abreu P, Hodge SE (1998) The power to detect linkage in complex disease by means of simple LOD-score analyses. *Am J Hum Genet* 63:870–879
- Hodge SE, Greenberg DA (1992) Sensitivity of lod scores to changes in diagnostic status. *Am J Hum Genet* 50:1053–1066
- Cordell HJ (2009) Detecting gene–gene interactions that underlie human diseases. *Nat Rev Genet* 10:392–404

18. Kooperberg C, Leblanc M, Dai JY, Rajapakse I (2009) Structures and assumptions: strategies to harness gene x gene and gene x environment interactions in GWAS. *Stat Sci* 24:472–488
19. Musani SK, Shriner D, Liu N, Feng R, Coffey CS, Yi N, Tiwari HK, Allison DB (2007) Detection of gene x gene interactions in genome-wide association studies of human population data. *Hum Hered* 63:67–84
20. Rodriguez-Murillo L, Subaran R, Stewart WC, Pramanik S, Marathe S, Barst RJ, Chung WK, Greenberg DA (2010) Novel loci interacting epistatically with bone morphogenetic protein receptor 2 cause familial pulmonary arterial hypertension. *J Heart Lung Transplant* 29:174–180
21. Balding DJ (2006) A tutorial on statistical methods for population association studies. *Nat Rev Genet* 7:781–791
22. Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 9:477–485
23. Lewontin RC (1995) The detection of linkage disequilibrium in molecular sequence data. *Genetics* 140:377–388
24. Jorde LB (2000) Linkage disequilibrium and the search for complex disease genes. *Genome Res* 10:1435–1444
25. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. *Nat Genet* 36:512–517
26. Sasieni PD (1997) From genotypes to genes: doubling the sample size. *Biometrics* 53:1253–1261
27. Weiss KM, Clark AG (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 18:19–24
28. Tysk C, Lindberg E, Jarnerot G, Floderus-Myrhed B (1988) Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. *Gut* 29:990–996
29. Thompson NP, Driscoll R, Pounder RE, Wakefield AJ (1996) Genetics versus environment in inflammatory bowel disease: results of a British twin study. *BMJ* 312:95–96
30. Orholm M, Binder V, Sorensen TI, Rasmussen LP, Kyvik KO (2000) Concordance of inflammatory bowel disease among Danish twins. Results of a nationwide study. *Scand J Gastroenterol* 35:1075–1081
31. Spehlmann ME, Begun AZ, Burghardt J, Lepage P, Raedler A, Schreiber S (2008) Epidemiology of inflammatory bowel disease in a German twin cohort: results of a nationwide study. *Inflamm Bowel Dis* 14:968–976
32. Pokorny RM, Hofmeister A, Galandiuk S, Dietz AB, Cohen ND, Neiberghs HL (1997) Crohn's disease and ulcerative colitis are associated with the DNA repair gene MLH1. *Ann Surg* 225:718–723. discussion 723–715
33. Franchimont D, Belaiche J, Louis E, Simon S, GrandBastien B, Gower-Rousseau C, Fontaine F, Delforge M (1997) Familial Crohn's disease: a study of 18 families. *Acta Gastroenterol Belg* 60:134–137
34. Polito JM II, Rees RC, Childs B, Mendeloff AI, Harris ML, Bayless TM (1996) Preliminary evidence for genetic anticipation in Crohn's disease. *Lancet* 347:798–800
35. Polito JM II, Childs B, Mellits ED, Tokayer AZ, Harris ML, Bayless TM (1996) Crohn's disease: influence of age at diagnosis on site and clinical type of disease. *Gastroenterology* 111:580–586
36. Hugot JP, Laurent-Puig P, Gower-Rousseau C, Olson JM, Lee JC, Beaugerie L, Naom I, Dupas JL, Van Gossum A, Orholm M, Bonaiti-Pellie C, Weissenbach J, Mathew CG, Lennard-Jones JE, Cortot A, Colombel JF, Thomas G (1996) Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* 379:821–823
37. Ohmen JD, Yang HY, Yamamoto KK, Zhao HY, Ma Y, Bentley LG, Huang Z, Gerwehr S, Pressman S, McElree C, Targan S, Rotter JI, Fischel-Ghodsian N (1996) Susceptibility locus for inflammatory bowel disease on chromosome 16 has a role in Crohn's disease, but not in ulcerative colitis. *Hum Mol Genet* 5:1679–1683
38. Parkes M, Satsangi J, Lathrop GM, Bell JI, Jewell DP (1996) Susceptibility loci in inflammatory bowel disease. *Lancet* 348:1588
39. Mirza MM, Lee J, Teare D, Hugot JP, Laurent-Puig P, Colombel JF, Hodgson SV, Thomas G, Easton DF, Lennard-Jones JE, Mathew CG (1998) Evidence of linkage of the inflammatory bowel disease susceptibility locus on chromosome 16 (IBD1) to ulcerative colitis. *J Med Genet* 35:218–221
40. Brant SR, Fu Y, Fields CT, Baltazar R, Ravenhill G, Pickles MR, Rohal PM, Mann J, Kirschner BS, Jabs EW, Bayless TM, Hanauer SB, Cho JH (1998) American families with Crohn's disease have strong evidence for linkage to chromosome 16 but not chromosome 12. *Gastroenterology* 115:1056–1061
41. Curran ME, Lau KF, Hampe J, Schreiber S, Bridger S, Macpherson AJ, Cardon LR, Sakul H, Harris TJ, Stokkers P, Van Deventer

- SJ, Mirza M, Raedler A, Krus W, Meckler U, Theuer D, Herrmann T, Gionchetti P, Lee J, Mathew C, Lennard-Jones J (1998) Genetic analysis of inflammatory bowel disease in a large European cohort supports linkage to chromosomes 12 and 16. *Gastroenterology* 115:1066–1071
42. Annese V, Latiano A, Bovio P, Forabosco P, Piepoli A, Lombardi G, Andreoli A, Astegiano M, Gionchetti P, Riegler G, Sturiniolo GC, Clementi M, Rappaport E, Fortina P, Devoto M, Gasparini P, Andriulli A (1999) Genetic analysis in Italian families with inflammatory bowel disease supports linkage to the IBD1 locus—a GISC study. *Eur J Hum Genet* 7:567–573
 43. Cavanaugh J, Consortium IBDIG (2001) International collaboration provides convincing linkage replication in complex disease through analysis of a large pooled data set: Crohn disease and chromosome 16. *Am J Hum Genet* 68:1165–1171
 44. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 250:1684–1689
 45. Narod SA, Amos C (1990) Estimating the power of linkage analysis in hereditary breast cancer. *Am J Hum Genet* 46:266–272
 46. Easton DF, Bishop DT, Ford D, Crockford GP (1993) Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. *Am J Hum Genet* 52:678–701
 47. Brown MA, Solomon E (1994) Towards cloning the familial breast-ovarian cancer gene on chromosome 17. *Curr Opin Genet Dev* 4:439–445
 48. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273:1516–1517
 49. Cuyvers E, Sleegers K (2016) Genetic variations underlying Alzheimer’s disease: evidence from genome-wide association studies and beyond. *Lancet Neurol* 15:857–868
 50. Karaderi T, Drong AW, Lindgren CM (2015) Insights into the genetic susceptibility to type 2 diabetes from genome-wide association studies of obesity-related traits. *Curr Diab Rep* 15:83
 51. Astradsson A, Olafsson E, Ludvigsson P, Bjorgvinsson H, Hauser WA (1998) Rolandic epilepsy: an incidence study in Iceland. *Epilepsia* 39:884–886
 52. Sidenvall R, Forsgren L, Heijbel J (1996) Prevalence and characteristics of epilepsy in children in northern Sweden. *Seizure* 5:139–146
 53. Heijbel J, Blom S, Rasmuson M (1975) Benign epilepsy of childhood with centrotemporal EEG foci: a genetic study. *Epilepsia* 16:285–293
 54. Bali B, Kull LL, Strug LJ, Clarke T, Murphy PL, Akman CI, Greenberg DA, Pal DK (2007) Autosomal dominant inheritance of centrotemporal sharp waves in rolandic epilepsy families. *Epilepsia* 48:2266–2272
 55. Echenne B, Cheminal R, Rivier F, Negre C, Touchon J, Billiard M (1992) Epileptic electroencephalographic abnormalities and developmental dysphasias: a study of 32 patients. *Brain Dev* 14:216–225
 56. Holtmann M, Becker K, Kentner-Figura B, Schmidt MH (2003) Increased frequency of rolandic spikes in ADHD children. *Epilepsia* 44:1241–1244
 57. Scabar A, Devescovi R, Blason L, Bravar L, Carozzi M (2006) Comorbidity of DCD and SLI: significance of epileptiform activity during sleep. *Child Care Health Dev* 32:733–739
 58. Strug LJ, Clarke T, Chiang T, Chien M, Baskurt Z, Li W, Dorfman R, Bali B, Wirrell E, Kugler SL, Mandelbaum DE, Wolf SM, McGoldrick P, Hardison H, Novotny EJ, Ju J, Greenberg DA, Russo JJ, Pal DK (2009) Centrottemporal sharp wave EEG trait in rolandic epilepsy maps to Elongator Protein Complex 4 (ELP4). *Eur J Hum Genet* 17:1171–1181
 59. Panjwani N, Wilson MD, Addis L, Crosbie J, Wirrell E, Auvin S, Caraballo RH, Kinali M, McCormick D, Oren C, Taylor J, Trounce J, Clarke T, Akman CI, Kugler SL, Mandelbaum DE, McGoldrick P, Wolf SM, Arnold P, Schachar R, Pal DK, Strug LJ (2016) A microRNA-328 binding site in PAX6 is associated with centrottemporal spikes of rolandic epilepsy. *Ann Clin Transl Neurol* 3:512–522
 60. Pal DK, Li W, Clarke T, Lieberman P, Strug LJ (2010) Pleiotropic effects of the 11p13 locus on developmental verbal dyspraxia and EEG centrottemporal sharp waves. *Genes Brain Behav* 9:1004–1012

INDEX

A

Affymetrix® 115, 119, 121, 122, 124, 142, 147
 Airway epithelial cells (AECs) 267–291
 Alzheimer’s disease (AD)..... 30, 62, 66, 67,
 94, 96, 99, 392
 Association analysisvi, 8, 136, 140, 382,
 384–389, 392, 393, 395

B

Bacteria 55, 58–60, 77–79, 81, 85
 Bisulfite conversion 233–251
 BRCA1 391, 392
 Brg1 368
 Brm 369

C

Caenorhabditis elegans 53–68
 Cancer genetics 368
 Cancer genomics 372
 Causal variants 4, 7, 11, 12, 318
 Cell characterization 20
 Cellular reprogramming 20
 Cluster generation..... 223, 229, 230
 Clustered regularly interspaced short palindromic repeats
 (CRISPRs) 29, 43, 44, 53, 55, 58
 Coding regions7, 163, 303,
 359, 373, 374, 392, 394
 Common diseases..... 6, 7, 146,
 304, 306, 307, 382, 389–393
 CpG 251
 Cq method 257–264
 Crohn’s disease (CD) 389–391

D

Data analysis 48, 49, 83, 115, 117,
 144, 223, 224, 234, 260, 263, 315, 333, 338
 Delta-delta C_q 258, 263, 264
 Diabetes v, vi, 3, 4, 7, 26, 30, 40,
 42, 86, 94, 99, 233, 257, 339, 354, 355, 362, 383
 DNA v, 3, 20, 42,
 56, 78, 91, 113, 152, 164, 176, 233, 268, 294,
 304, 355, 371, 387
 DNA substitutions 357
 Dual-luciferase assay.....305, 310, 313, 315

E

Electroporation296–298, 300
 Epigenetics24, 27, 42, 175, 234, 307, 369

F

Ferroportin 1 (SLC40A1) 356
 Firefly luciferase305, 306, 313, 315
 Full-length cDNA176, 183, 199–220
 Functional variant(s)29, 303–318

G

GenABEL 115, 118–120, 125–144
 Gene editing..... 43, 44, 49, 55
 Gene expression profilingvi, 27, 30, 374, 376
 Gene knockout268, 270, 282, 285
 Gene silencing 293–301
 Gene therapy 39–49
 Genetic epidemiology 371, 376, 389
 Genetic heterogeneity384, 391, 392, 395
 Genetic risk score 30, 41, 113, 323, 334
 Genetic screens.....57, 63, 64, 66, 67
 Genetics v, 3, 17,
 40, 54, 78, 91, 113, 163, 233, 267, 303, 323,
 354, 368, 381
 Genome-wide association (GWA)v, vi, 4,
 27, 113–148, 324
 Genome-wide association studies (GWAS) 6–8,
 12, 17, 27, 41, 303, 304, 323–339, 381, 382,
 384, 386, 388, 392, 393
 Genomic screens 59
 Green fluorescence protein (GFP) 54, 55,
 57, 60–62, 68

H

Hemochromatosis vi, 353–362
 Hemojuvelin (HJV) 356, 359, 362
 Hepatocellular carcinoma94–97
 Hepsidin (HAMP)356
 HFE 355, 357–359,
 361, 362
 Human complex disease v, 3, 6,
 12, 30, 324, 338, 382–384, 386, 389
 Human diseases v, 18, 24–27,
 40–42, 48, 68, 86, 99, 113, 223, 381, 395

I

Illumina® 5, 6, 79–81,
84, 115, 117, 119, 121, 122, 124, 142, 147, 152,
164, 166, 170, 171, 178, 190, 194, 197, 198,
201, 217, 220, 223, 226, 228, 232, 234–236,
239, 241–243, 245, 247–251, 325
Induced pluripotent stem cells (iPSCs) 17–31, 46
Iron absorption 355
Iron overload 354–356, 362

L

Lentiviruses 21, 43
Library preparations 80, 152,
153, 164–170, 172, 181, 183, 219, 220,
226–228, 230, 232, 235
Linkage analysis vi, 324, 357,
359, 362, 381–395
Linkage mapping 356, 358
Long insert whole genome sequencing 151–154,
156–159
Long noncoding RNAs (lncRNAs) v, 91–103, 176

M

Mathematical modeling 325, 326, 332
Methylation vi, 27, 42, 251
Microbiome v, 77–87
MicroRNAs (miRNAs) 21, 24,
42, 43, 46, 92, 94, 96, 99, 234, 264, 310
mRNAs 21, 23, 24,
43, 92, 93, 99, 175, 177, 220, 257, 293, 304, 310

N

Next generation sequencing (NGS) 3–12,
40, 49, 161, 173, 199, 223–232, 258, 362, 368,
370, 372, 381
Noncoding RNAs 7, 12, 49, 99

P

Personalized medicine 335
Pharmacogenomics 40, 45
Phenotyping 27, 28, 31,
326, 339, 383, 390, 391, 393
Precision medicine 39–41, 48
Primary cells 297, 300

Q

Quantitative PCR (qPCR) 81, 85,
229, 231, 258, 262–264
qRT-PCR 271, 279, 289
Quantitative traits 116, 125,
130, 138, 140, 145, 307, 327, 335, 339

R

Rare cancers 367–376
Rare diseases 3, 7, 303, 376
Recombination fraction 383, 385
Regression analysis 332, 336
Renilla luciferase 305, 306,
310, 313–315, 318
Reverse transcription (RT) 200, 202, 206,
207, 220, 258, 260–262, 298
Reverse transcription polymerase chain reaction
(RT-PCR) 298
RNA interference (RNAi) vi, 53,
55, 57, 58, 60–64, 66–68, 301
R programming software 118–120,
126–128, 136, 141, 144, 146

S

Sequencing-by-synthesis (SBS) 5, 6, 152, 223
Short insert whole genome sequencing 152, 153,
157, 159
Short interfering RNA (siRNA) 43, 58,
61, 92, 293–300
Single-cell RNA-sequencing 199
Single-cell sequencing 9
Single-cell transcriptomics 9, 10,
27, 94, 177, 220
Small cell carcinoma of the ovary, hypercalcemic type
(SCCOHT) 377
SMARCA2 369, 373
SMARCA4 377
SNP annotation 115, 144, 145
Somatic mosaicism 9
16S rDNA 78–80, 84
16S rRNA 84
Statistics 249

T

Targeted amplicon sequencing 78, 81, 85
Transfection 12, 22, 23,
44, 45, 293, 295–301, 305, 307, 310, 311,
313–315, 317, 318, 357
Transferrin receptor 2 (TFR2) 356, 359, 361, 362

U

Ultralow abundance 175, 181

W

Whole exome sequencing (WES) 7, 9,
11, 163–173, 373, 374
Whole genome sequencing (WGS) v, 4, 6,
7, 9, 11, 12, 31, 57, 151–161, 164, 304, 338, 355,
382, 395