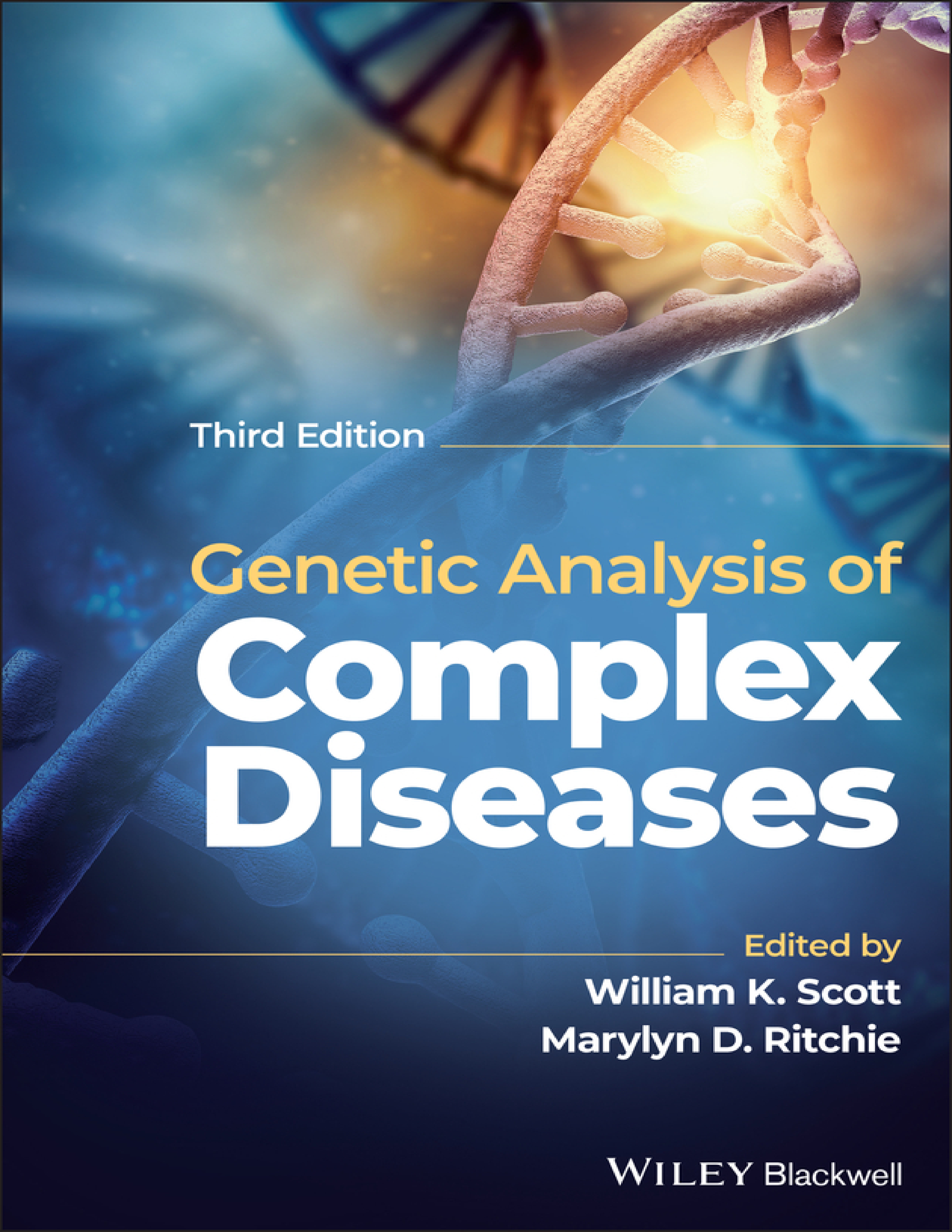

Third Edition

Genetic Analysis of Complex Diseases

Edited by
William K. Scott
Marylyn D. Ritchie

WILEY Blackwell

Table of Contents

- Cover
- Title Page
- Copyright Page
- List of Contributors
- Foreword
- 1 Designing a Study for Identifying Genes in Complex Traits
  - Introduction
  - Components of a Disease Gene Discovery Study
  - Keys to a Successful Study
  - References
- 2 Basic Concepts in Genetics
  - Introduction
  - Historical Contributions
  - DNA, Genes, and Chromosomes
  - Genes, Mitosis, and Meiosis
  - Inheritance Patterns in Mendelian Disease
  - Genetic Changes Associated with Disease/Trait Phenotypes
  - Susceptibility Versus Causative Genes
  - Summary
  - References
- 3 Determining the Genetic Component of a Disease
  - Introduction
  - Study Design

- Approaches to Determining the Genetic Component of a Disease
- Summary
- References

4 Study Design for Genetic Studies

- Introduction
- Selecting a Study Population
- Family-Based Studies (Linkage)
- Family-Based Studies (Association)
- Cohort Studies
- Cross-Sectional Studies
- Case–Control Studies
- Other Study Designs
- Biobanks
- Other Biobanks
- Biospecimens for Biobanks
- Summary
- References

5 Responsible Conduct of Research in Genetic Studies

- Introduction
- Research Regulations and Genetics Research
- Addressing Pertinent ELSI in Genetic Research
- Practical Methods for Efficient High-Quality Genetic Research Services
- References

6 Linkage Analysis

- Disease Gene Discovery
- Ability to Detect Linkage

Real World Example of LOD Score Calculation and Interpretation
Disease Gene Localization
Multipoint Analysis
Effects of Misspecified Model Parameters in LOD Score Analysis
Impact of Incorrect Disease Allele Frequency
Impact of Incorrect Mode of Inheritance
Impact of Incorrect Disease Penetrance
Impact of Incorrect Marker Allele Frequency
Control of Scoring Errors
Genetic Heterogeneity
Practical Approach for Model-Based Linkage Analysis of Complex Traits
Nonparametric Linkage Analysis
Identity by State and Identity by Descent
Methods for Nonparametric Linkage Analysis
Tests for Linkage Using Affected Sibling Pairs (ASP)
Tests Based on Identity by Descent in ASPs
Multipoint Affected Sib-Pair Methods
Methods Incorporating Affected Relative Pairs
NPL Analysis
Fitting Population Parameters
Power Analysis and Experimental Design Considerations for Qualitative Traits
Examples of Sib-Pair Methods for Mapping Complex Traits
Mapping Quantitative Traits
Measuring Genetic Effects in Quantitative Traits

- Study Design for Quantitative Trait Linkage Analysis
- Variance Components Linkage Analysis
- Nonparametric Methods
- The Future
- References

7 Data Management

- Developing a Data Organization Strategy
- Database Management System (DBMS) and Structured Query Language (SQL)
- Database Implementation
- Other Tools for Data Management and Manipulation
- Conclusion
- References

8 Linkage Disequilibrium and Association Analysis

- Introduction
- Linkage Disequilibrium
- Summary
- References

9 Genome-Wide Association Studies

- Introduction
- Design
- Data Analysis
- Conclusion
- References

10 Bioinformatics of Human Genetic Disease Studies

- Introduction
- Common Threads Genome Analysis

Processing and Analysis of Genomic Data
Bioinformatics Resources
References
11 Complex Genetic Interactions/Data Mining/Dimensionality Reduction
Human Diseases Are Complex
Complexity of Biological Systems
Statistical and Mathematical Concepts of Complex Genetic Models
Analytic Approaches to the Detection of Complex Interactions
Conclusion
References
12 Sample Size, Power, and Data Simulation
Introduction
Sample Size and Power
Power Calculations and Simulation
Power Studies for Association Analysis
Power Simulations for Linkage Analysis
Summary
References
Index
End User License Agreement

List of Tables

Chapter 2
Table 2.1 Useful applications of Hardy–Weinberg theory.

Table 2.2 Differences between meiosis and mitosis.

Table 2.3 Hallmarks of Mendelian inheritance patterns of different types.

Table 2.4 Salient features of human repeat expansion diseases.

Chapter 3

Table 3.1 Association between sex of offspring and risk of expansion in fra...

Table 3.2 The association between disease concordance rates in twins and di...

Table 3.3 Adoptive studies and disease etiology.

Table 3.4 Familial correlation and heritability estimates of pulmonary func...

Chapter 5

Table 5.1 Resources regarding human subjects genetics research and regulati...

Table 5.2 Information on the Genetic Information Nondiscrimination Act (GIN...

Chapter 6

Table 6.1 Example development of genetic map using four linked Loci*A, B, C*,...

Table 6.2 LOD scores for pedigrees in Examples 1, 2, and 3.

Table 6.3 Number of phase-known, fully informative meioses needed to detect...

Table 6.4 Two-point linkage analysis for Alzheimer disease and D19S246.

Table 6.5 Impact of misspecifying disease allele frequency on LOD score ana...

Table 6.6 Impact of misspecifying mode of inheritance on LOD score analysis...

Table 6.7 Impact of misspecifying disease penetrance on LOD score analysis.

Table 6.8 Impact of misspecifying marker allele frequencies on LOD score an...

Table 6.9 Expected percentage of affected pairs showing 0, 1, or 2 alleles ...

Table 6.10 Analysis of IBS sharing probabilities for two siblings at a mark...

Table 6.11 Results of simple Sibpair tests on IDDM data.

Table 6.12 Expected LOD scores for mapping testicular cancer susceptibility...

Table 6.13 Familial Correlations in Blood Pressure.

Chapter 7

Table 7.1 Software and web resources.

Chapter 8

Table 8.1 Measures of allelic association for alleles *A* and *B* at different l...

Table 8.2 Case–control association studies: *APOE-4* allele and AD.

Table 8.3 Summary of epidemiological measures.

Table 8.4 2 × 2 Contingency table for case–control analysis.

Table 8.5 Example of population stratification[a].

Table 8.6 Multiallelic TDT: T_{mhet}[a].

Table 8.7 Transmission disequilibrium test and diabetes[a] .

Chapter 11

Table 11.1 Penetrance values for combinations of genotypes from two SNPs ex...

Table 11.2 Penetrance values for combinations of genotypes from two SNPs ex...

Chapter 12

Table 12.1 The four possible outcomes of an experiment.

Table 12.2 Practical considerations when determining sample size.

Table 12.3 Number of sibships and nuclear families in variance component ...

List of Illustrations

Chapter 1

Figure 1.1 Steps in a Mendelian disease gene discovery (positional cloning) ...

Figure 1.2 Study cycle for a complex trait gene identification study.

Figure 1.3 Components of a complex disease study and expertise needed to con...

Chapter 2

Figure 2.1 Principles of Mendel’s first law of segregation of heritable char...

Figure 2.2 Principles of Mendel’s second law of independent assortment with ...

Figure 2.3 The DNA double helix is packaged and condensed in several differe...

Figure 2.4 The genetic code and abbreviations for amino acids.

Figure 2.5 Central dogma of genetics: DNA → RNA → protein.

Figure 2.6 A G-banded human male karyotype.

Figure 2.7 The myotonic dystrophy (DM) and insulin receptor (INSR) genes are...

Figure 2.8 Genetic results of crossing over: (a) no crossover: A and B remai...

Figure 2.9 Genes that are on the same chromosome (syntenic) may be unlinked ...

Figure 2.10 Pedigrees consistent with (a) autosomal dominant inheritance, (b...

Figure 2.11 Single base pair changes in exon 4 of APOE define the 2, 3, and ...

Chapter 3

Figure 3.1 Ascertainment schemes for genetic analysis.

Figure 3.2 Example of ascertainment bias in genetic analysis when ascertaini...

Figure 3.3 Correlation of age of onset among siblings affected with Alzheime...

Chapter 6

Figure 6.1 (a) Pedigree in which a rare, fully penetrant autosomal dominant ...

Figure 6.2 (a) Pedigree in which a rare, fully penetrant autosomal dominant ...

Figure 6.3 Pedigree for Example 3 for calculation of LOD score for linkageph...

Figure 6.4 Pedigree examples demonstrating families that are informative and...

Figure 6.5 Pedigree on the left shows unordered genotypes underneath pedigre...

Figure 6.6 The most likely location for the disease gene is in the 7-cM inte...

Figure 6.7 Example of multipoint linkage analysis in the presence of genetic...

Figure 6.8 Examples of identity by state and identity by descent. See text f...

Figure 6.9 Example showing the inclusion of additional family members (here ...

Figure 6.10 Power calculations for MLS sibpair analysis.

Chapter 7

Figure 7.1 Data normalization example for dataset with information mapping S...

Chapter 8

Figure 8.1 Decay of allelic association for recombination fractions (Θ) of 0...

Figure 8.2 An example of haplotype blocks in a 100 kb region on chromosome 1...

Figure 8.3 Transmitted and non-transmitted alleles in a family triad.

Figure 8.4 Example of scoring a TDT family.

Chapter 9

Figure 9.1 Example of raw genotype intensity clusters for a single SNP. Imag...

Figure 9.2 A flowchart overview of the entire GWAS QC process. Each topic is...

Figure 9.3 Example of a principal components analysis plot – visualizing the...

Figure 9.4 Example of two Q-Q plots for a GWAS dataset. The image on the lef...

Chapter 10

Figure 10.1 Tsne visualization of methylation array data clusters for pediat...

Figure 10.2 View from the UCSC genome browser of a segment of the *CTNNA2* gen...

Chapter 11

Figure 11.1 Genotype models. Given a single biallelic variant, there are thr...

Genetic Analysis of Complex Diseases

Third Edition

Edited by William K. Scott and Marylyn D. Ritchie

WILEY Blackwell

This third edition first published 2022

Edition History
2nd edition © 2006 by John Wiley & Sons, Inc. All rights reserved.

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
9600 Garsington Road, Oxford, OX4 2DQ, UK
For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Scott, William K., 1970– editor. | Ritchie, Marylyn DeRiggi, 1977– editor.
Title: Genetic analysis of complex diseases / edited by William K. Scott and Marylyn D. Ritchie.
Description: Third edition. | Hoboken, NJ : Wiley-Blackwell, 2022. | Preceded by Genetic analysis of complex diseases / [edited by] Jonathan L. Haines, Margaret Pericak-Vance. 2nd ed. c2006. | Includes bibliographical references and index.
Identifiers: LCCN 2021009896 (print) | LCCN 2021009897 (ebook) | ISBN 9781118123911 (paperback) | ISBN 9781119104087 (adobe pdf) | ISBN 9781119104070 (epub)
Subjects: MESH: Genetic Diseases, Inborn–genetics | Disease–genetics | Chromosome Mapping–methods | Genetic Predisposition to Disease | Research Design | Genetic Research
Classification: LCC RB155 (print) | LCC RB155 (ebook) | NLM QZ 50 | DDC 616/.042–dc23
LC record available at https://lccn.loc.gov/2021009896
LC ebook record available at https://lccn.loc.gov/2021009897

Cover Design: Wiley
Cover Image: © ESB Professional/Shutterstock

List of Contributors

Susan H. Blanton
Dr. John T. Macdonald Foundation
Department of Human Genetics
University of Miami Miller School of Medicine
Miami, FL, USA

Adam Buchanan
Genomic Medicine Institute
Geisinger
Danville, PA, USA

William S. Bush
Department of Population and Quantitative Health Sciences
Case Western Reserve University
Cleveland, OH, USA

Ren-Hua Chung
Institute of Population Health Sciences
Division of Biostatistics and Bioinformatics
National Health Research Institutes (Taiwan)
Hsinchu, Taiwan

Dana C. Crawford
Department of Population and Quantitative Health Sciences
Case Western Reserve University
Cleveland, OH, USA

Abigail Deppen
InformedDNA
St Petersburg, FL, USA

Logan Dumitrescu
Department of Neurology
Vanderbilt University
Nashville, TN, USA

Kayla Fourzali
University of Miami Miller School of Medicine
Miami, FL, USA

Susan Estabrooks Hahn
Genomic Services
Quest Diagnostics
North Andover, MA, USA

Jonathan L. Haines
Department of Population and Quantitative Health Sciences
Case Western Reserve University
Cleveland, OH, USA

Dale J. Hedges
Center for Applied Bioinformatics
St. Jude Children's Research Hospital
Memphis, TN, USA

Elizabeth Heise
Clinical Genetics Program
GeneDX, Inc
Gaithersburg, MD, USA

Allison Ashley Koch
Duke Molecular Physiology Institute
Duke University Medical Center
Durham, NC, USA

Eden R. Martin
Dr. John T. Macdonald Foundation
Department of Human Genetics
University of Miami Miller School of Medicine
Miami, FL, USA

Jacob L. McCauley
Dr. John T. Macdonald Foundation
Department of Human Genetics
University of Miami Miller School of Medicine
Miami, FL, USA

Sarah A. Pendergrass
Human Genetics
Genentech
San Francisco, CA, USA

Margaret A. Pericak-Vance
Dr. John T. Macdonald Foundation
Department of Human Genetics
University of Miami Miller School of Medicine
Miami, FL, USA

Evadnie Rampersaud
Center for Applied Bioinformatics
St. Jude Children's Research Hospital
Memphis, TN, USA

Marylyn D. Ritchie
Department of Genetics
Perelman School of Medicine at the University of Pennsylvania
Philadelphia, PA, USA

William K. Scott
Dr. John T. Macdonald Foundation
Department of Human Genetics
University of Miami Miller School of Medicine
Miami, FL, USA

Stephen D. Turner
Signature Science
LLC
Charlottesville, VA, USA

Shefali Setia Verma
Department of Pathology and Laboratory Medicine
Perelman School of Medicine at the
University of Pennsylvania
Philadelphia, PA, USA

Yogasudha Veturi
Department of Genetics
Perelman School of Medicine at the University of Pennsylvania
Philadelphia, PA, USA

Chantelle Wolpert
Physician Assistant Program
Thomas Jefferson University
Philadelphia, PA, USA

Foreword

This book grew from our four-day NIH-sponsored course, which, for 20 years, was focused on providing an overview and guide to the design and execution of human genetic mapping studies for these common (and genetically complex) diseases, melding the genomic technology with the statistical rigor needed to apply and interpret the results. When we developed the concept for the first edition of this book in 1996, the Human Genome Project was just reaching full speed, combining continual breakthroughs in DNA gene mapping and sequencing technology with emerging applications to human disease to shed the first light on the organization of the human genome and the variations that cause disease. The first applications of the Human Genome Project data were to find the location, and ultimately the causative mutations, for rare Mendelian inherited diseases. It was dogma then that the genetic architecture of common diseases was beyond our reach, based on the naïve belief that Mendelian disease represented how genetic variation impacted disease. However, we soon demonstrated, with the discovery that multiple apolipoprotein E (*APOE)* alleles had differing and strong effects on the risk of Alzheimer disease, that these technologies and approaches could be adapted to illuminate the genetic underpinnings of common diseases.

The rapid advances in both DNA technology and statistical methodology demanded that a significant update to the book was needed, with the second edition of the book in 2006. By this point the blood and protein markers of the 1970s had been surpassed by the restriction fragment length polymorphisms (RFLPs) of the 1980s, the microsatellite repeats of the 1990s, and the single nucleotide polymorphisms (SNPs, of which RFLPs are a subset) for the past 20 years. Naturally, the analyses of these data also advanced from early mainframe applications of genetic linkage analysis in small numbers of families, to PC-powered analyses of thousands of cases and controls for association.

In the past 15 years since that second edition, increasingly dense SNP arrays and whole exome or whole genome sequencing have created new horizons for dissecting complex diseases. In addition, the explosion of other “omics” data, particularly gene expression data, provide biological context for the discovered DNA variations, adding biological interpretation as a critical element of genetic studies.

With all these advances, it became apparent that a new edition of this book was warranted, and new and fresh perspectives were needed. Thus, we turned over the editing of this new edition to two of our brilliant younger colleagues, who have been active in both developing and applying methods at the forefront of genetics and genomics. While the inclusion of genome-wide association studies, integration of genomic data, and data mining are new, the breadth of the book in describing the overall process of designing and executing successful projects remains.

Finally, we fondly acknowledge the continuing impact of our mentor, Dr. P. Michael Conneally, who inspired both of us to inquire, question, investigate, and solve, the often difficult, constantly emerging human genetic puzzles. He encouraged us to help educate researchers, physician-scientists, and physicians in the complex nature of genetic studies. He wrote the forward for the first two editions, and although he passed away in 2017, his legacy remains in our work and the work of our trainees and collaborators.

We are immensely grateful to Bill and Marylyn for taking on this important task and developing this excellent third edition of the book.

Jonathan L. Haines, PhD
Margaret A. Pericak-Vance, PhD

1
Designing a Study for Identifying Genes in Complex Traits

William K. Scott[1], Marylyn D. Ritchie[2], Jonathan L. Haines[3], and Margaret A. Pericak-Vance[1]

[1] Dr. John T. Macdonald Foundation Department of Human Genetics, University of Miami Miller School of Medicine, Miami, FL, USA

[2] Department of Genetics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA

[3] Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, USA

Introduction

Disease gene discovery in humans has a long history, predating even the identification of DNA as the genetic molecule (Watson and Crick 1953) and the determination of the number of human chromosomes (Ford and Hamerton 1956; Tjio and Levan 1956). In fact, as early as the 1930s some simple statistical methods for the analysis of genetic data had been developed (Bernstein 1931; Fisher 1935a,b). However, these methods were severely limited in their application (more on basic concepts of genetics in Chapter 2). Not only were genetic markers lacking (the ABO blood type was one of the few that had been described), but these methods were restricted to small, two to three generation pedigrees. Any calculations were performed by hand, of course, making analysis laborious.

There were two hurdles to overcome before human disease gene discovery would become routine. First, appropriate statistical methods were lacking, as were ways of automating the calculations. Second, sufficient genetic markers to cover the human genome needed to be identified. Morton (1955), building on the work of Haldane and Smith (1947) and Wald (1947), described the use of maximum likelihood approaches in a sequential test for linkage between two loci. He used the term “LOD score” (for logarithm of the odds of linkage) for his test. This score is the basis for most modern genetic linkage analyses and represents a milestone in human disease gene discovery. However, the complex calculations had to be done by hand, severely limiting the use of this approach. Elston and Stewart (1971) described a general approach for calculating the likelihood of any non-consanguineous pedigree. This algorithm was extended by Lange and Elston (1975) to include pedigrees of arbitrary complexity. Soon thereafter, the first general-purpose computer program for linkage in humans, LIPED (Ott 1974), was described. Thus, the first of the two major hurdles was overcome.

By the mid-1970s there were 40–50 red cell antigen and serum protein polymorphisms available as genetic markers. A few markers could be arranged into initial linkage groups, but these markers covered only approximately 5–15% of the human genome. In addition to this limited coverage, genotyping these polymorphisms was labor intensive, time consuming, and often quite technically demanding. This remaining hurdle was crossed with the description of restriction fragment length polymorphisms (RFLPs) by Botstein et al. (1980). Not only were these markers easier to genotype in a standard manner, but they were frequent in the genome, covering the remaining 85–95% of the genome for the first time.

With these tools in place, the field of human disease gene discovery blossomed. The first successful disease gene linkage using RFLPs was reported (Gusella et al. 1983), localizing the Huntington disease gene to chromosome 4p. This discovery marked the beginning of disease gene identification through the *positional cloning* approach. Early successes using positional cloning were for diseases inherited in Mendelian fashion: autosomal dominant, autosomal recessive, or X-linked. Although confounding factors such as genetic heterogeneity, variable penetrance, and phenocopies might exist for single-gene or Mendelian traits, it is generally possible with a known genetic model to determine the best and most efficient approach to identifying the responsible gene. The success of these tools is apparent since by mid-2017 over 3350 single-gene disorders had at least one causative genetic variant identified (OMIM, accessed May 2017 at http://omim.org).

However, the inheritance patterns for traits such as the common form of Alzheimer’s disease, multiple sclerosis, and non-insulin-dependent diabetes (to name a few) do not fit any simple genetic explanation, making it far more difficult to determine the best approach to identifying the unknown underlying effect. In addition to the confounding factors involved in single-gene disorders, such as genetic heterogeneity and phenocopies, gene–gene and gene–environment interactions must be considered when a complex trait is dissected. However, the tools that enabled efficient mapping of Mendelian trait loci through positional cloning were not as effective in dissecting these more complex traits. New statistical tools, study designs, and genotyping technologies were needed to perform large-scale analysis of genetic factors underlying these complex traits. As these technologies were developed, a new approach to complex disease gene identification via genome-wide association studies (GWAS) was enabled. The shift to this approach was predicted by a seminal perspective published by Risch and Merikangas (1996), in which they showed that large-scale case–control analyses of complex traits would be a powerful and efficient method of identifying alleles underlying complex traits, once genotyping technology allowed the cost-effective determination of a dense map of genetic markers. The first GWAS was published in 2005 (Klein et al. 2005), identifying the association of variation in the *CFH* gene with age-related macular degeneration. This was simultaneously confirmed using alternate study designs (Edwards et al. 2005; Haines et al. 2005) proving that GWAS worked, allowing this new era of complex disease genetics to begin in earnest.

With the dawn of the GWAS era, a corresponding shift in the prevailing hypotheses for these studies occurred. No longer were studies solely searching for one or a few rare mutations in a single gene that cause a rare and devastating disease. Studies of common complex diseases were searching for multiple alterations in one or more genes acting alone or in concert to increase or decrease the risk of developing a trait. Early GWAS tended to test the “common disease-common variant” (CDCV) hypothesis: the risk for common diseases, across ethnic groups, arises from evolutionarily old variants that have had substantial time to spread throughout the human population. Many studies successfully identified thousands of variants associated with the risk of complex diseases. An interactive catalog of these variants is maintained by the National Human Genome Research Institute and the European Molecular Biology Laboratory at http://www.ebi.ac.uk/gwas. Despite these successes, many studies testing the CDCV hypothesis failed to explain all the heritable variation in the risk of the complex traits under study – a phenomenon termed “missing heritability” (Manolio et al. 2009). One explanation for this was that the effect of rare variants was not well studied by early GWAS – an alternative hypothesis termed the “common disease-rare variant” (CDRV) hypothesis. This hypothesis suggests that risk of common complex diseases arises from a larger number of rare variants in one or more genes, perhaps occurring more recently.

As was the case with common variants and the exploration of the CDCV hypothesis being enabled by GWAS approaches and high-throughput genotyping technology, exploration of the CDRV hypothesis was enabled by advances in high-throughput sequencing technology and accompanying statistical analysis methods. Initial screens of coding-sequence variants in Mendelian traits via whole-exome sequencing (WES) were published by Ng et al. (2009, 2010) and Choi et al. (2009), demonstrating that in some cases, disease gene mapping could skip the positional cloning strategy and proceed directly to evaluating segregation of mutations in families. This proof of principle has been used to justify this approach for testing the CDRV hypothesis in complex traits but has been met with mixed success. A successful example is the recent analysis of 50 000 individuals in the MyCode Community Health Initiative successfully identified rare variants underlying cardiovascular traits and lipid levels (Dewey et al. 2016). The rapid and continuing decrease in whole-genome sequencing (WGS) costs suggests that within a few years, it will be possible (and perhaps commonplace) to test the CDRV hypothesis using WGS in large sample sizes – essentially performing genome-wide association for common and rare variants with direct genotype determination via sequencing.

Study design, laboratory methods, and analytic approaches differ by trait type (Mendelian or complex) and hypothesis being tested (rare disease-rare variant, Mendelian positional cloning; CDCV [GWAS]; CDRV [WES or WGS and individual variant or set-based association]). These approaches are described in the following sections.

Components of a Disease Gene Discovery Study

Each genetically complex trait has its own peculiarities that require special attention. However, a guiding paradigm can be applied to most conditions. Originally, the general approach that was used for Mendelian single-gene disorders was *positional cloning*. With the completion of the human genome reference sequence, cloning was no longer a necessary step – and therefore this general approach is better described as *disease gene discovery*. The classical approach (Figure 1.1) follows a generally linear series of events: defining the phenotype, identifying multi-case families, collecting blood samples, genotyping markers, analyzing data for initial disease gene localization, refining the initial localization to define the minimum candidate region, and then sequencing genes within this region to find the causative mutation(s).

In contrast to the classical approach, the current approaches to finding genes for common and genetically complex traits are not linear, and many steps are works in progress, subject to further defining, refining, or replacement by subsequent steps. Figure 1.2 illustrates the stepwise and recursive nature of the components of a complex trait study. Each step has its own key factors that must be considered, and for complex traits, the order and emphasis of these steps on the approach will vary from study to study. This fact is underappreciated and contrasts strongly with the classical disease gene discovery approach. Indeed, many of the difficulties reconciling discordant studies of the same complex trait arise from study-specific decisions made in the approach.

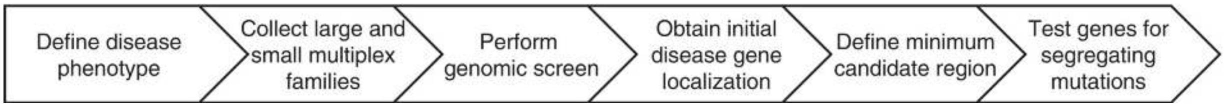

Figure 1.1 Steps in a Mendelian disease gene discovery (positional cloning) study.

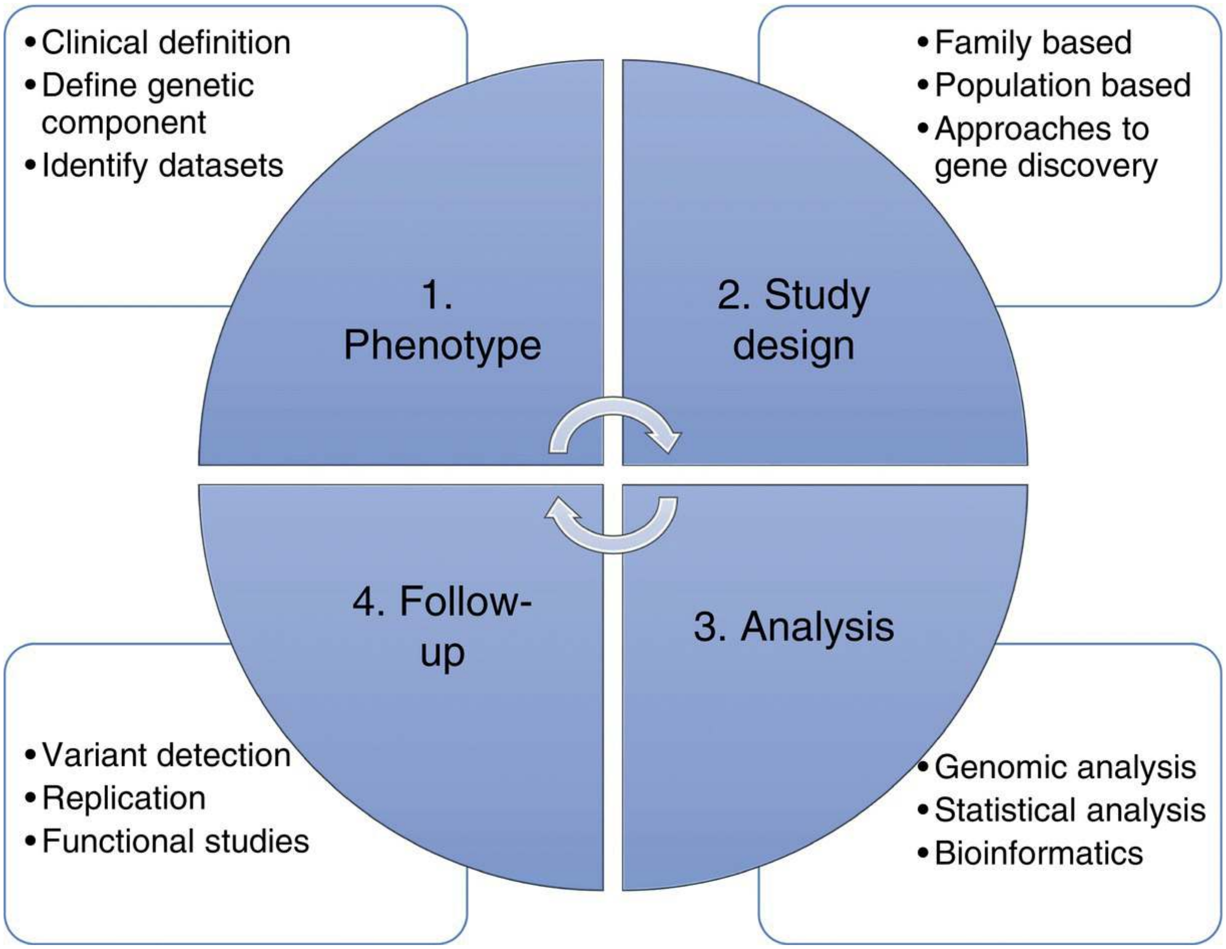

Figure 1.2 Study cycle for a complex trait gene identification study.

This section discusses the steps in Figure 1.2, providing an overview of each component and a guide to the chapter(s) providing more detail on these points.

Define Disease Phenotype

The first step in any disease gene discovery process is to know what phenotype is being studied. This may sound obvious, but specifying the exact measures that will be used to reliably and validly determine the phenotype is often overlooked in the rush to move forward. There are three aspects that need to be considered: clinical definition, determining that a trait has a genetic component, and identification of datasets that can be studied.

Clinical Definition

It is not enough to define a trait in binary terms, such as the presence or absence of Huntington’s disease or diabetes. In Huntington’s disease, for example, there can be wide variation in the symptoms, with some only psychological or very mild motor disturbances detectable by expert examination, and the age at which these symptoms begin is similarly variable. In diabetes, there are distinct subtypes (insulin-dependent diabetes mellitus and non-insulin-dependent diabetes mellitus) as well as variable age at onset. Additionally, blood glucose levels (a quantitative trait) are strongly associated with diabetes (a qualitative trait) and could be used as a surrogate measure or endophenotype. One critical role of the clinician in study design is to assess the various diagnostic procedures and tools and determine which ones best define a consistent phenotype. Additionally, dissecting genetically complex diseases usually requires large datasets to supply enough power to unravel genetic effects. For this reason, participant ascertainment often extends to multiple sites. It is critical for multi-site studies to establish consensus diagnostic procedures and criteria and apply them consistently across sites. For example, the establishment of a consensus diagnostic scheme (McKhann et al. 1984) played an important role in a successful complex disease linkage study in late-onset familial Alzheimer’s disease (Pericak-Vance et al. 1991) and subsequent identification of the association of Alzheimer’s disease and common