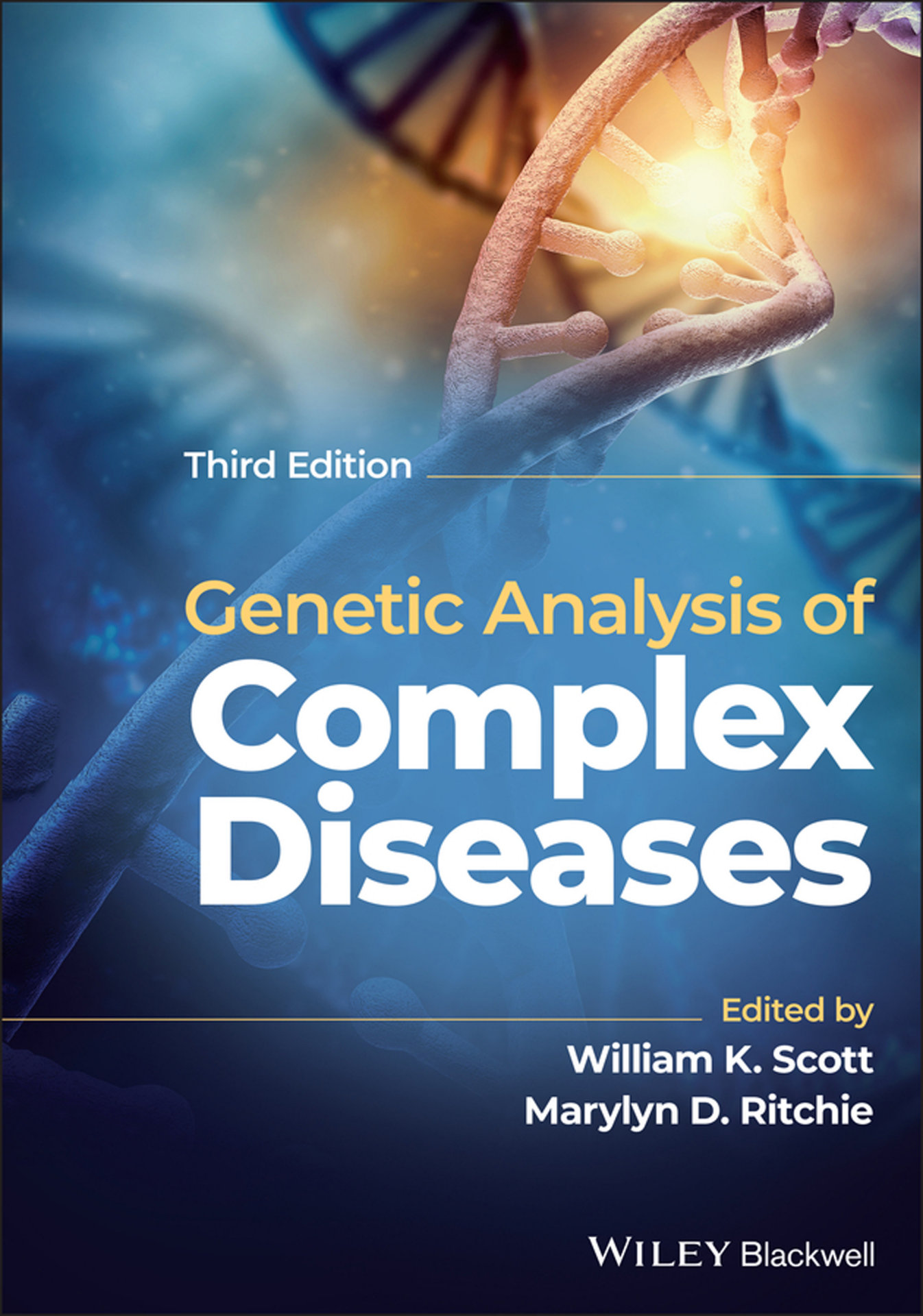
Third Edition
Genetic Analysis of
Complex
Diseases
Edited by
William K. Scott
Marylyn D. Ritchie
WILEY Blackwell

Genetic Analysis of Complex Diseases

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

Set in 9.5/12.5pt STIXTwoText by Straive, Pondicherry, India

10 9 8 7 6 5 4 3 2 1

Contents

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

Genetic Analysis of Complex Diseases, Third Edition. Edited by William K. Scott and Marylyn D. Ritchie.

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

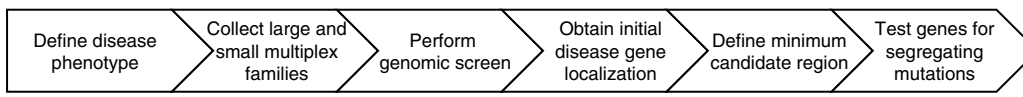

Figure 1.1 Steps in a Mendelian disease gene discovery (positional cloning) study.

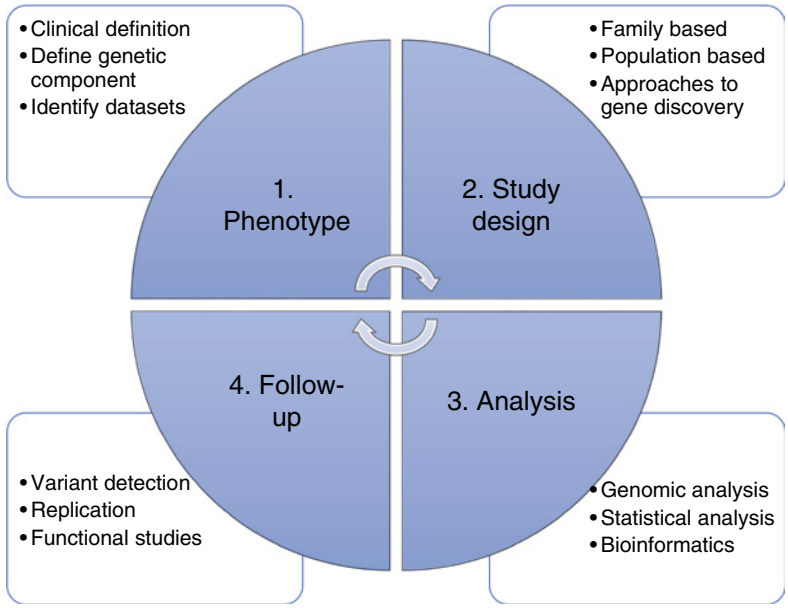

Figure 1.2 Study cycle for a complex trait gene identification study.

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.

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 variation in the *APOE* gene (Corder et al. 1993; Corder et al. 1994).

The phenotype assignment must be done in a rigorously consistent fashion. Even a small rate of phenotype error might alter analytic results – in some cases leading to false-positive results and in others to false-negative results. Thus, which data will be used to assign the trait status must be carefully determined. Must detailed clinical records of an examination specifically addressing the phenotype be obtained and reviewed for consistency on every participant? Is the self-report of a participant or a participant's relative sufficient? Is a note documenting a diagnosis (but no examination findings) from a medical record adequate? Or is direct examination of every participant using a standardized research protocol required? Additionally, investigators must consider whether to collect additional biomarker data (e.g. antibody titers, protein assays) or clinical tests (e.g. electroencephalogram, electrocardiogram, magnetic resonance imaging) that might correlate with the trait of interest. The goal of the phenotyping protocol is to standardize procedures, minimize error in determining the phenotype, and maximize the power of the dataset to detect genes underlying the trait.

Determining that a Trait Has a Genetic Component

It is critical that as much as possible be known about the genetic basis of a complex trait prior to determine the most appropriate study design for gene identification. That a trait "runs in families" is insufficient evidence, since this phenomenon can occur for several reasons other than shared genetic susceptibility, including shared environmental exposure and biased ascertainment. As outlined in Chapter 3, there are numerous lines of evidence that can be examined, including family studies, segregation analysis, twin studies, adoption studies, heritability studies, and population-based risks to relatives of probands (the initially identified individual with disease). For most traits being contemplated, some such data already exist in the literature. A thorough review of this literature may provide most of the necessary information and point out any missing data. The data may not only indicate the strength of the genetic effect on the trait but also give some indication of the underlying genetic model. For example, there may be obvious evidence of a single "major" gene, such as in Huntington's disease, or multiple genes interacting in complex ways, such as in multiple sclerosis (Sadovnick et al. 1996).

Identification of Datasets

It is helpful to identify early on what potential datasets exist or can be collected. Do large families exist or are most cases apparently sporadic? Are large cohort or case–control studies available? Are there repositories of multiplex families with associated clinical data available? Are there existing clinical networks or large specialty clinics available? Is the necessary phenotype data available in a biobank linked to an existing electronic health record? The answers to these questions determine what study designs are feasible for the trait under study, as discussed in Chapters 3 and 4.

Develop Study Design

Developing your study design and delineating the phenotype are not independent steps. Review of the available data may indicate that a trait as originally defined has little or no evidence of a genetic component. However, there may be strong evidence that a subset of the trait is strongly genetic. For

example, there had for many years been debate about the role of genetics in Alzheimer's disease. Over time it became increasingly clear that a subset of individuals with the onset of Alzheimer's disease before age 65 existed and strongly clustered in families with apparent autosomal dominant inheritance. Within each of these families, Alzheimer's disease appeared to be caused by a single gene. By restricting sample collection and genetic analysis to these types of families, three genes (*APP*, *PSEN1*, and *PSEN2*) were identified with mutations causing early-onset Alzheimer's disease (Goate et al. 1991; Levy-Lahad et al. 1995; Rogaev et al. 1995; Sherrington et al. 1995).

The exact approach to the disease gene discovery process should be outlined as completely as possible before the project gets underway. With the clinical phenotype in hand, it is possible to determine the best strategy for defining what type of dataset to collect. Participant recruitment is perhaps the longest and most labor-intensive step in the entire process. It is imperative that the enrollment of participants (particularly if studying multiple members of the same family) proceeds with careful consideration of the wishes and norms of the participating individuals, families, and communities. The rights of individuals to participate or refuse participation should receive careful consideration, and the informed consent process should provide adequate explanation of the study and answer any questions, and, critically, confidentiality must be carefully protected. These issues are outlined in detail in Chapter 5.

Determination of the study design (case–control, cohort, case series, family-based) is based on the characteristics of the phenotype, the estimated genetic model, and the research objective. For example, the existence of large families with apparent Mendelian segregation suggests that a single major gene could be detected, and a family-based study would be appropriate. A phenotype with weaker estimated heritability, a pattern of recurrence risks suggesting many genes of small effect, and little familial aggregation would suggest that a case–control study design is most feasible. The process of selecting a study design to answer a research question is reviewed in Chapter 4.

It is also important to have some sense of the sample size required to identify the genes being sought. When pedigree structures are already available in family-based studies of single-gene disorders, power is easily calculated with high confidence for specific genetic models using computer simulation programs. For complex traits, however, genetic models are not as easily specified in advance, and computer simulations often must consider a range of parameter values for the genetic model to describe the power across several competing alternatives. Chapter 12 provides an overview of the available approaches and tools for sample size, power estimation, and genetic simulations.

Family-Based Studies

Family-based studies include large extended families, smaller multi-case families (often affected sibpair or other affected relative pairs), and discordant sibpair studies. Depending on family structure and number of individuals collected, these families may be used in linkage analyses (as discussed in Chapter 6) or association studies (Chapter 8). Depending on the genetic architecture of the trait and the frequency of the disease-associated alleles being sought, this design may offer increased power over population-based designs.

Population-Based Studies

Several types of observational designs may be considered for population-based studies, including case-series, case–control, and longitudinal cohort designs. The possible sampling frames for these types of studies include simple random samples of a defined geographical area, clinic- or hospital-based samples, convenience samples such as voluntary registries or biobanks, or hybrids of these (e.g. health-system-based biobanks linked to longitudinal electronic health records). These designs

became much more frequent with the advent of high-throughput genotyping technologies, which enabled the efficient study of very large samples of unrelated individuals through GWAS (Chapter 9), an approach with substantially greater power than a similarly sized family-based study.

Approaches for Gene Discovery

There are two general, but not mutually exclusive, ways to approach gene discovery for complex traits. The first is to take a genome-wide screening approach. Genomic screening can aim to identify areas of *genetic linkage* in family-based designs (Chapter 6) or areas of *association* in either family- or population-based designs (Chapters 8 and 9). A good genomic screen will attempt to cover the entire human genome using markers evenly spaced across the genome. Current high-throughput genotyping technologies enable genotyping of hundreds of thousands to millions of single nucleotide polymorphisms in a rapid, inexpensive manner for use in linkage or association studies. More recently, high-throughput sequencing technology has been used to screen the entire coding sequence of the genome (WES) or the entire genome (WGS) for trait-associated variants, without first conducting genome-wide linkage or association studies. As sequencing costs continue to decline, a shift to "genotyping by sequencing" is likely, in which results from WGS might be used to conduct a genome-wide screen and follow-up in a single molecular experiment. These same high-throughput genotyping and sequencing technologies allow large-scale examination of gene expression (through gene expression microarrays or RNA-Seq) and epigenetic changes (through methylation arrays or Methyl-Seq) in trait-relevant tissues. The results of such experiments are often used in conjunction with genome-wide screens to identify high-priority candidate genes for follow-up studies. These technologies and their application to genomic studies are discussed in Chapter 10.

In contrast to the genomic screening approach, a directed screening approach may be used. This approach, sometimes termed a "candidate-gene" approach, focuses on an area of the genome selected for examination based on prior information. The additional information could come from many sources, including results from a previous genome-wide screen, results from gene expression studies, genes suggested by pathophysiology, or candidate genes identified in model systems. For example, multiple sclerosis is an autoimmune disease in which the myelin sheaths around nerves are attacked and often destroyed. This information suggests that certain genes, such as the human leukocyte antigen genes, T-cell receptor genes, and the myelin basic protein gene, are prime candidates for analysis. The strength and weakness of this approach arise from the confidence in the role of these genes. If the evidence is strong that a direct role is played, only a few such genes may need to be tested to find a trait-associated variant. If the evidence is more circumstantial, then many genes may have equal justification for being studied, and not much is gained over conducting a genome-wide screen. Such studies are now most often conducted as follow-up of prior genomic screens or other hypothesis-generating experiments.

Analysis

Genomic Analysis

Generally, genome-wide genotyping or sequencing is the first analytic step. Such studies may use newly collected blood samples or stored blood samples (or extracted DNA or RNA) made available by a biorepository. Depending on the goal of the study and its design, genome-wide genotyping, sequencing, gene expression, or epigenetic analysis may be performed on these samples. Some studies may be able to re-use stored genotype or sequence data available from public repositories (such as dbGaP [https://www.ncbi.nlm.nih.gov/gap] or the European Genome-phenome Archive

[https://www.ebi.ac.uk/ega/home]) or from prior studies of the sample being used. The technologies and approaches to these molecular experiments are covered in Chapter 10. In each case, it is important to formulate a quality control plan to detect potential laboratory errors such as sample switches, failed genotyping probes, sequencing errors, and batch effects. When possible, coordinating laboratory analysis with initial analytic quality control is optimal for finding and correcting such errors. If archived genomic data are being used, careful review of the initial quality control protocols and further checks (when possible) in the subsequent analysis is recommended.

Statistical Analysis

The analysis of genetic and phenotypic data for a complex trait is multifaceted and depends on the research question, study design, genomic data available, and phenotypic characteristics. Methods to analyze these data are under constant development, and new approaches are continuously being released. Therefore, the analytic strategy for a genomic study must be reviewed periodically and revised if necessary to take advantage of newly developed approaches. Depending on the study design, the analytic plan may include linkage analysis (Chapter 6) in families or association studies in families or population samples (Chapters 8 and 9). These approaches are not mutually exclusive – a design may start with a linkage analysis of large families followed by association analysis within regions of linkage. Similarly, other multi-stage studies conduct a GWAS of individual SNPs (Chapter 9) and then incorporate gene–gene and gene–environment interactions to identify additional genetic loci. Additionally, "data mining" approaches may be applied to these datasets to extract even more genetic information using data reduction techniques, set-based tests, and pathway analyses. These more complex analyses are discussed in detail in Chapter 11.

Bioinformatics

The large amount of information generated by any genomic study of a complex trait requires careful attention to quality control, efficient and secure storage, and compliance with data-sharing requirements and privacy protections. These activities require a well-designed and secure database system. Such systems have evolved over time from text files to relational databases, to large-scale "data warehouses." Such datasets also require large-scale processing power with ample attached storage to facilitate linkage and association studies. High-throughput sequencing in particular requires a large amount of storage and computational power for genome alignment (or assembly) and base calling. For multi-site studies, these resources may need to be accessible from multiple locations, requiring levels of access and security depending on the role on the study and need to access other sites' information. In addition to maintaining local resources for a study, a bioinformatics team also must be familiar with many different public sources of genomic data (e.g. UCSC and Ensembl browsers, ENCODE databases, sequence repositories, dbGaP) and be able to submit results to public repositories for sharing with the wider research community. These issues are discussed in more detail in Chapter 7.

Follow-up

Variant Detection

Once a single gene (or region) is implicated by a screen (linkage or association), it is necessary to examine it for potentially functional variations that might explain the linkage or association signal. For positional cloning efforts, this generally consisted of sequencing the minimum candidate region and identifying mutations that segregated with the trait in families. For complex traits, this effort is more difficult, and the variant being sought may be a more common, yet functional,

polymorphism. Several strategies, including haplotype analysis, conditional analysis, and exhaustive sequencing, may be used in this case. The analyses required for such efforts are discussed in Chapters 8 and 9. However, statistical analysis of a single dataset only goes so far to establish a trait-associated variant. Additional studies, including replication in independent datasets and functional studies in cellular and animal models, may be required to ultimately determine if a variant influences the biology underlying the complex trait.

Replication

The literature on most complex traits is at this point littered with initial reports of allelic or genotypic associations that cannot be replicated at all (or are replicated in a small minority of studies). Reproducibility of findings in independent samples is a critical characteristic most investigators seek when weighing the evidence for a trait-associated variant. Because of this, most studies (particularly those seeking government or foundation funding) now include a plan for replication of findings in a second dataset. These replication datasets should be independent of the initial finding (e.g. do not overlap with the discovery dataset) and be assessed in similar fashion (e.g. phenotype definitions agree, ascertainment is similar, genetic analysis is comparable). This does not mean that the datasets must be from the same population – indeed, demonstrating replication across populations (e.g. European, Asian, and African) for a common complex trait locus may add strength to the study. However, for rare variants, cross-population replication might be more difficult (due to population-specific alleles); for such studies, replication in a second sample from the sample population would be desirable.

Functional Studies

While most disease gene discovery efforts have claimed success based on finding variants that segregate with traits in pedigrees or polymorphisms significantly associated with the trait in population samples, this is, strictly, not sufficient evidence. More conclusive is evidence arising from biological systems (e.g. cultured cells, animal models, or human blood and tissue samples) that the trait can be either induced by introduction of the allele or ameliorated by blocking the action of the allele. In genetically complex traits, where the responsible variation may be a common polymorphism, it is even more critical that such evidence be found before success is declared.

Tests in biological systems can be of several types. Perhaps the most common is to test the action of the gene in a model organism, such as mouse, zebrafish, or fruit fly. With transgenic models, the proposed trait-associated variant is introduced into the germline of the organism and the resulting offspring are examined for evidence of the abnormal phenotype. With knockout models, the action of the gene in question is eliminated and the offspring are examined for evidence of an abnormal phenotype. Similar experiments can be performed in cultured cells, where the introduction of the variant (or gene knockout) is easier. However, finding the appropriate cell line and determining the appropriate cellular phenotype corresponding to the trait may be difficult. Recent advances in generating relevant cellular models have utilized inducible pluripotent stem cell (iPSC) technology, by which cells (blood, fibroblast) from an individual with a phenotype and genotype of interest can be reprogrammed and differentiated to a cell type of interest (such as neuron or retinal pigment epithelium). Such cells might be closer to the affected tissue type and have more recognizable phenotypes due to the genetic variant under study. A further advance incorporates gene editing technology (e.g. CRISPR/Cas9) into the approach, whereby an established iPSC line can be edited to introduce (or correct) a variant of interest. Such an approach eliminates the need to draw a sample from a person known to carry a variant of interest and allows examination of isogenic cell lines with and without the variant for phenotypic changes. These approaches are rapidly evolving,

and frequently revised sources, such as *Current Protocols in Human Genetics*, should be consulted for the latest details on functional studies using these approaches.

Keys to a Successful Study

Foster Interaction of Necessary Expertise

To appropriately carry out any disease gene discovery study, one must use techniques from five different areas of expertise (Figure 1.3). These areas are clinical evaluation, molecular genetics, statistical genetics, bioinformatics, and epidemiology. The first provides the necessary diagnostic and participant recruitment skills needed to define the phenotype and help collect samples and data. The second provides genotyping, sequencing, and functional analysis skills necessary to help locate and identify the genes and variants of interest and evaluate their functional consequences. The third provides the statistical and analytical framework for the proper design of the study and the analysis of the generated data. The fourth provides computational and algorithmic expertise for the processing, storage, and dissemination of large-scale datasets. And the fifth provides expertise to incorporate environmental variables and apply results at the population level.

The initial focus of gene discovery on single-gene disorders resulted in a linear approach (Figure 1.1) that could be implemented by a single investigator with expertise in one of these areas, with periodic consultation with colleagues from other disciplines as needed. Complex traits require a multidisciplinary approach that is not easily implemented by a single investigator, and given differences in genetic architecture, available samples, and research questions, different approaches (and thus different teams) may need to be formed for each trait. Thus, experts in each of these

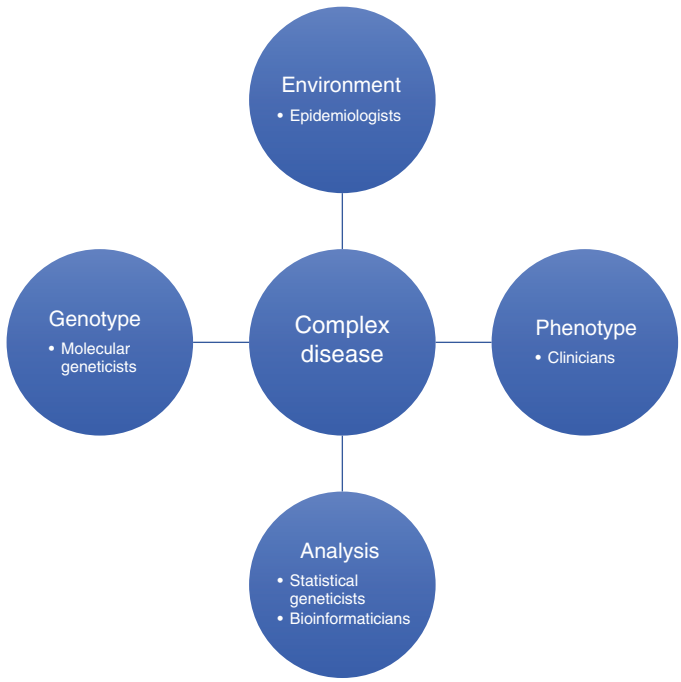

Figure 1.3 Components of a complex disease study and expertise needed to contribute.