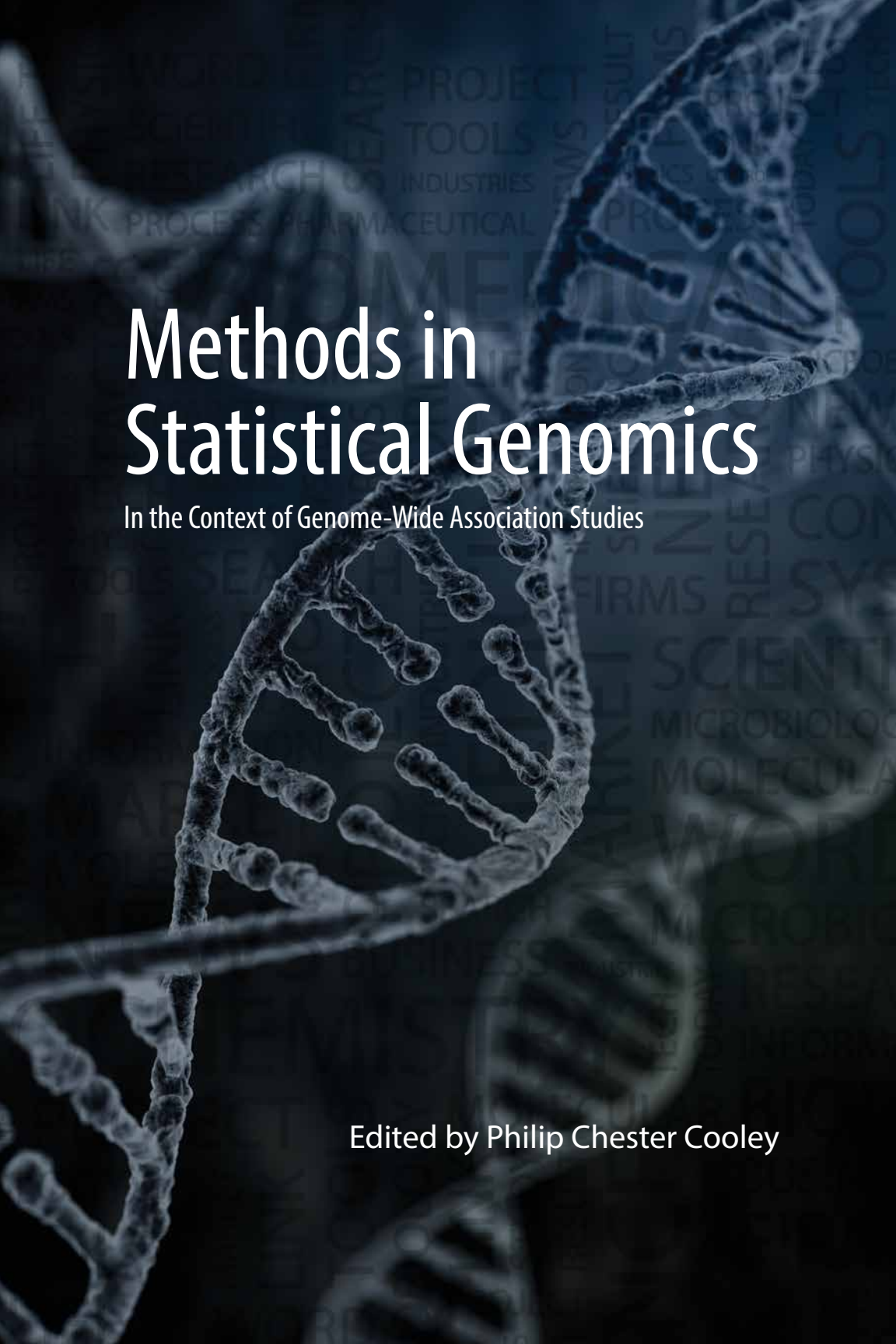


Methods in Statistical Genomics

In the Context of Genome-Wide Association Studies

Edited by Philip Chester Cooley

Methods in Statistical Genomics: In the Context of Genome-Wide Association Studies

Edited by
Philip Chester Cooley

RTI Press

RTI INTERNATIONAL

Library of Congress Control Number: 2016949391

ISBN 978-1-934831-16-8
(refers to print version)

RTI Press publication No. BK-0016-1608
https://doi.org/10.3768/rtipress.2016.bk.0016.1608
www.rti.org/rtipress

Cover design: John Theilgard

The RTI Press mission is to disseminate information about RTI research, analytic tools, and technical expertise to a national and international audience. RTI Press publications are peer-reviewed by at least two independent substantive experts and one or more Press editors.

RTI International is an independent, nonprofit research institute dedicated to improving the human condition. We combine scientific rigor and technical expertise in social and laboratory sciences, engineering, and international development to deliver solutions to the critical needs of clients worldwide.

This publication is part of the RTI Press Book series.

RTI International
3040 East Cornwallis Road, PO Box 12194
Research Triangle Park, NC 27709-2194, USA
rtipress@rti.org
www.rti.org

Contents

CHAPTER 1

Overview of Chapters

Philip Chester Cooley

Introduction

The objective of this book is to describe procedures for analyzing genome-wide association studies (GWAS). Some of the material is unpublished and contains commentary and unpublished research; other material (Chapters 4 through 7) has been published previously. Each previously published chapter investigates a different genomics model, but all focus on identifying the strengths and limitations of various statistical procedures that have been applied to different GWAS scenarios.

The distinction between genotype and phenotype was initially presented by the Danish botanist, plant physiologist, and geneticist Wilhelm Johannsen in a book he published in 1905, *The Elements of Heredity*. He distinguished between the genotype of the organism (it is hereditary) and the ways in which its heredity is demonstrated in phenotypes, or physical characteristics. This distinction was an outgrowth of Johannsen's experiments concerning heritable variation in plants.[1]

Today, it is understood that the process leading from genes to proteins that ultimately establish phenotypes is complex. Most proteins are the products of multiple genes. Whether a protein is an enzyme, receptor, or hormone, it functions in a specific environment that includes external factors like temperature, rainfall, the amount of sunlight available, and nutrition, as well as internal factors that can include other hormones, enzymes, and other proteins.

Further, biochemical pathways are not always linear; they can have multiple positive and negative feedback loops and may involve multiple steps and the products of hundreds of genes. In summary, the evolutionary forces producing a phenotype may often involve many genes and can be influenced by a variety of specific environmental factors.

The Human Genome Project

The Human Genome Project (HGP) was an international scientific project with the goals of determining the sequence of chemical base pairs that make up human DNA and of identifying all of the physical and functional genes of the human genome. The HGP produced the first complete sequences of individual human genomes. As of 2012, thousands of human genomes had been completely sequenced, and many more had been mapped at lower levels of resolution. The resulting data have been used worldwide in biomedical science, anthropology, forensics, and other branches of science. With the mapping of the human genome near completion, researchers expected that subsequent genomic studies would lead to advances in our understanding of human evolution, and advances in many subfields of biology, particularly the diagnosis and treatment of many diseases.

To that end, researchers have worked to identify genes that constitute biomarkers using a combination of high-throughput experimental and bioinformatics approaches; nevertheless, the identification of biological functions of the protein and RNA products of DNA has only just begun. Recent results suggest that most of the vast quantities of noncoding DNA within the genome have biochemical activities that include regulating gene expression, organizing chromosome architecture, and producing signals that control epigenetic inheritance.[2]

A major aim of the HGP was to determine the functions of genes. Researchers believed that once the complete genome sequence was developed, interpreting the sequence by comparing the intermediate messenger RNA and protein products would be straightforward and ultimately would identify the genetic factors that influence important phenotypes such as predisposition to certain diseases. The simple rationale behind GWAS is that if certain genetic variations are more frequent in persons with a given disease, the variations are said to be "associated" with the disease. The associated genetic variations serve as pointers to regions of the human genome that may be involved in causing the disease.

Genome-Wide Association Studies

GWAS compare the DNA of two groups of participants: subjects with the phenotype of interest (cases, or persons with a particular disease) and similar subjects without the phenotype (controls). Each subject provides a sample of DNA, from which millions of genetic variants are read using single

polymorphism (SNP) arrays. If one type of the variant (one allele, i.e., the "wild-type" allele) is more frequent in people with the disease, the SNP is said to be associated with the disease. The associated SNPs are then considered to mark a region of the human genome that influences the risk of the phenotype. Also, in contrast to methods which specifically test one or a few genetic regions, GWAS investigate the entire genome. The approach is therefore said to be non–candidate driven, in contrast to gene-specific candidate-driven studies. GWAS identify tag SNPs, which are defined as representative SNPs in a region of the genome with high linkage disequilibrium and other variants in DNA associated with a disease. Tag SNPs in isolation cannot specify which genes cause the phenotype.

The first successful GWAS investigated age-related macular degeneration and was published in 2005.[3] This study found two SNPs that had significantly altered allele frequency when compared with healthy controls. As of 2015, *The Catalog of Published Genome-Wide Association Studies* contained more than 2,141 catalog entries, 1,856 publications and 12,874 implicated SNPs.[4] Prior to the introduction of GWAS, the major method of investigation was via genetic linkage studies in families. This approach was useful for identifying single-gene disorders, many of which appear in the comprehensive compendium of human genes and genetic phenotypes, the Online Mendelian Inheritance in Man (OMIM) database.[5]

However, for both common and complex diseases, the results of genetic linkage studies have been hard to reproduce.[6,7] In contrast, GWAS seek to identify whether the allele of a genetic variant is found more often than expected in individuals with the phenotype of interest. The statistical methods used in GWAS are based on traditional approaches, and early calculations of statistical power indicated that GWAS could be better than linkage studies at detecting weak genetic effects.[8]

In addition to a simple conceptual framework, the proliferation of GWAS has also been driven by improvements in sequencing methods, reduced computational costs, and the advent of biobanks, which are repositories of human genetic material that greatly reduce the cost and difficulty of collecting sufficient numbers of biological specimens for study.[9,10] The development of rapid genome-level sequencing techniques also permits researchers to assess methods to mine this information to identify genetic associations with disease and ultimately determine the biological basis of disease patterns. Knowing the coding sequences of every nucleotide in an organism has permitted researchers to study the collective influence of all genes simultaneously and their role in structuring organism traits, including specific diseases.

With improving genotyping technologies and the exponentially growing number of available markers, case-control GWAS have become a key tool for investigating complex diseases. To accommodate GWAS methods, researchers have developed new procedures to ensure data quality, interpret GWAS findings, and provide computationally tractable approaches when performing hundreds of thousands of individual tests.

The promise of GWAS was anticipated in many quarters of the scientific community.[11] A 2007 fact sheet released by the National Human Genome Research Institute, in the early days of GWAS, raised expectations that personalized medicine, including individual risk prediction, disease prevention, and specific treatment, was just around the corner. "With the first GWAS published in 2005, … health professionals will be able to use such tools to provide patients with individualized information about their risks of developing certain diseases … to tailor prevention programs to each person's unique genetic makeup … to select the treatments most likely to be effective and least likely to cause adverse reactions…."[12] However, a number of critics of GWAS argue that these expectations have not been met.[13]

GWAS: Useful or Misleading?

The overly high expectations were created in part because the early GWAS success predicting age-related macular degeneration with a complement factor H polymorphism was extreme, with an odds ratio (OR) of approximately 7.* In contrast, most GWAS implemented after that success involved variants conferring small effects, indicated by ORs slightly larger than 1.

Although medical science is still far from the GWAS-based personalized medicine promised in the National Human Genome Research Institute 2007 fact sheet, at least three important considerations fuel legitimate hope that genetics will become integral to a form of medicine more specifically tailored to individual patients.[11] First, important discoveries have already changed medical practice and resulted in medical policy codes for some treatments. For example, in the field of pharmacogenomics, researchers have already begun using genetic testing to determine patients' dosage of warfarin.

Second, genetic interaction studies are starting to provide useful data. For example, researchers discovered that high-density lipoprotein cholesterol levels

* The OR is a measure of is a measure of association between an exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

(HDL-C), one of the most important risk factors for coronary heart disease, are significantly influenced by the interplay of multiple genes linked to GWAS and involved gene-gene interaction effects.[14]

Third, GWAS are based on common variants (i.e., tag SNPs) that are frequently in linkage disequilibrium with the actual causative variant, which in turn may be associated with larger effect sizes than the common variant included in the GWAS. For instance, fine mapping of loci associated with low-density lipoprotein cholesterol (LDL-C) identified a rare nonsynonymous variant gene that explained 5 times more of the contributed variance than the initial GWAS finding. In this context, whole genome sequence data has the potential to be a more accurate and powerful tool than SNPs to elucidate the relationship between genetics and (common) diseases. However, even high-resolution genetic variation will only explain a fraction of the heritability of human diseases and traits. Thus, we are still searching for potential uses for genetics in medical science beyond using simple genetics with gene-gene and gene-environment interactions and identifying epigenetic effects as important but complex targets.

Missing Heritability

Heritability is a genetic measure that identifies the observable differences in a trait due to genetic factors between individuals within a population. Factors including genetics, environment and random chance can all contribute to the variation between individuals in their phenotypes.[15] Heritability is a dynamic measurement that identifies the fraction of phenotype *variability* that can be attributed to genetic *variation*. The term "missing heritability" refers to the low percentage of information about the overall genetic component and risk of common diseases gleaned from GWAS. Common variants account for only a small proportion of genetic components, and the missing heritability lies in the huge class of rare genetic variants that GWAS do not see. Variants that are primary drivers of disease are relatively rare in the human population.

Variants confer risk of disease, and natural selection acts against variants so that they do not become too prevalent. Therefore, the issue of so-called missing heritability becomes moot. We did not interrogate the whole genome; we interrogated the common variants that pass through the filter of natural selection. Many diseases, but not all, will involve rare variants not detected by GWAS.[16]

Autoimmune diseases may be an exception to this thinking. Some variants that are major risk factors appear more commonly in the general population;

for example, variants selected in response to infectious agents that have consequences for autoimmune diseases.[17]

The inability of GWAS to replicate markers in this and other instances was a major concern, and in a set of *New England Journal of Medicine* articles three authors offer alternative opinions regarding the progress that GWAS methods have made to date. Hirshhorn argued that the main goal of these studies is not the prediction of genetic risks but rather the discovery of biological pathways underlying polygenic diseases and traits.[18] Goldstein countered that the genetic burden of common diseases must be carried mainly by large numbers of rare variants.[19] He also suggested that most GWAS produced too many associations with very small overall contributions toward explaining disease risk—too many to provide any useful biologic insights. Kraft and colleagues support the notion that the GWAS approach predicts that "many, rather than few, variant risk alleles are responsible for the majority of the inherited risk of each common disease. It is possible that these initial GWAS have identified only the strongest associations, with many more genes still to be identified."[20]

Organization of the Book

Chapter 2: Genome Wide Association Data: Where Are the Standards?

Chapter 2 is a previously unpublished manuscript based on an internal study directed by Philip Cooley. Study participants included Huaqin Pan (RTI), Paul S. Levy (deceased), Maureen K. Bunger (formerly of RTI), and Laxminarayana Ganapathi (RTI Health Solutions).

We began our explorations into GWAS by applying statistical methods to real genetic (SNP) data. We used standard statistical approaches that appear in both GWAS and non-GWAS literature. The main objective of this venture was to see if we could assess the performance of different statistical methods used in a GWAS context. There have been many notable instances in which GWAS provided inconsistent results and there have been a number of studies have not been replicable. This has led to a question about the validity of the GWAS approach. For example, in 2008, the online listing *A Catalog of Published Genome-Wide Association Studies* identified a total of 8 amyotrophic lateral sclerosis (ALS) GWAS.[4] Each of these 8 studies identified candidate ALS markers (or alternatively no markers), but none of the studies could replicate the results of the other studies. As of 2015, the number of ALS GWAS has grown to 16 studies that use a variety of international data and have identified

some possible genes linking ALS to genetic causes. By and large, no consistent markers have been universally accepted as ALS genetic markers.

We obtained the SNP data for the initial ALS study appearing in the literature[21] and attempted to replicate some of the published results in order to test statistical methods in genomics.[21] We identified seven distinct methods that have been or could be applied in GWAS studies. We also used the method that Schymick et al. used to obtain their results. At the time we performed these analyses, we were unaware of any comprehensive studies that compared the performance of the different methods in the context of GWAS, and we wanted to determine whether standards could be developed. Using a previously conducted study would allow us to assess the performance of specific statistical methods by using the study results as a yardstick by which to measure the accuracy of our results.

What we learned from this effort is that either many of the algorithms used in the literature in a GWAS context assume an additive gene model or are agnostic with respect to the form of inheritance. We also documented the inability of the ALS studies to replicate results—part of the problem with reproducing results is that ALS studies rely on having ALS patients as a study population, which means that the original sample sizes were relatively small due to circumstance. We were able to replicate the Schymick et al. study results using the method they chose to measure associations—the classic epidemiology case-control method that uses the Pearson χ^2 test to text how likely it is that an observed distribution of data fits with the distribution that is expected if the variables are independent. However, our assessment indicated that the reported results depended to an unknown degree on the statistical method used to make the predictions. This suggested that the algorithms selected could influence predictive outcome and further demonstrated the need for developing GWAS standards.

In summary, the absence of both methodological standards and a process for evaluating the statistical methods used in predicting associations between genes and phenotypes suggested to us that we could examine these missing elements more effectively by using simulation methods. Accordingly, we created simulated data that were linked to known outcomes which therefore constituted a "truth set." The simulated data could be analyzed using different statistical methods, and we could assess each method's predictive properties, which could potentially reveal some or all of this missing information.

Chapter 3: Creating the Synthetic Gene Data

Chapter 3 describes the generic data generation process we used for the research described in Chapters 4 through 7. We used our simulated data in an effort to exploit a process with a known outcome to identify those traits that affect GWAS outcomes. These investigations took the form of (1) generating a database of synthetic genes that incorporated a number of dynamic properties that were varied for the explicit purpose of developing a "truth set" (i.e., a database of known outcomes); (2) testing a number of statistical models and competing algorithms that were developed to predict associations; and (3) creating a compendium of outcomes that linked gene properties to statistical model performance. This virtual gene resource enabled the power performance of different single-gene statistical methods to be measured and recorded. Consequently, this chapter describes the simulation model that generated the synthetic SNP data for all subsequent assessments represented in this book.

Chapter 4: Genetic Inheritance and Genome-Wide Association Statistical Test Performance Using Simulated Data

Chapter 4 is based on a study that was published previously.[22] This chapter focuses on single-gene models and the statistical methods that effectively predict associations in a GWAS context. Here, we demonstrate that the choice of a statistical method can affect the power profiles of GWAS predictions. The initial step in a GWAS is to apply univariate statistical tests for each SNP marker in the data set. Applying the tests is methodologically straightforward. SNP-based tests are used to assess the likelihood of an association. In the simulation-generated gene data, the probability of the occurrence of the phenotype is regulated by an exogenously specified risk value that is a function of the genotype. This allows the "strength" of the genotype-phenotype signal to be controlled and power outcomes to be counted. Standard methods (e.g., χ^2 tests, logistic regressions) are commonly used in single-locus tests. In general, the GWAS approach is a brute force method that scans the entire genome to determine which genes demonstrate an association using a stringent threshold level to compensate for the problem of multiple comparisons. The problem of multiple test comparisons arises because as the number of tests (hypotheses) increases, the likelihood of witnessing a rare event increases. Often the mode of genetic inheritance (MOI) is assumed to be additive, which implies that the allele causes the phenotype risk rather than the genotype.

Our simulation studies confirmed the results of others[23] that the gene model or MOI was a major influence on statistical power. This was no surprise and it is well known that associations involving recessive MOI SNPs are much more difficult to detect than other MOI types. Gene traits that influence prediction accuracy had also been reported in other studies.[24,25] They demonstrated how the phenotype MOI assumption was a major influence on association prediction accuracy.

We compared the power profiles of GWAS using a number of statistical methods, including two that combine MOI-specific methods into multiple test measures. Because most GWAS investigations have not determined the specific gene model operating, many assume an additive model. Given that most models cited in OMIM are either dominant or recessive gene models, we investigated composite methods that did not make an additive assumption; rather, they used three component tests with three distinct MOI properties (dominant, recessive, and additive). We then compared the performance (from a statistical power perspective) of the composite tests as contrasted with methods that either assume a specific MOI gene model or are agnostic with respect to MOI. By combining recessive, additive, and dominant individual tests, we determined that if the MOI is not known, then a composite test is more likely to make a correct association prediction. In that sense, it constitutes a more powerful test and could have significant advantages with respect to single test procedures.

Our findings did not provide a specific answer about which statistical method is best. The best method depends on the MOI gene model associated with the phenotype (diagnosis) in question and how common the traits are that associate with the phenotype. However, our results do indicate that the common additive assumption that the MOI of the locus is associated with the diagnosis can have adverse consequences. It indicated that researchers should consider a multitest procedure that combines the results of individual MOI-based core tests as a statistical method for conducting the initial screen in a GWAS. The process for combining the core tests into a single operational test can occur in a number of ways. We identify two: the Bonferroni procedure and the MAX procedure, each of which produces very similar statistical power profiles.[26,27] This was a surprising result because the Bonferroni procedure was based on combining χ^2 tests and assumed that the tests were mutually independent when they clearly are not, whereas the MAX procedure used

normal distribution tests and adjusted for the covariance properties between individual tests.

In summary, the focus of this chapter is on single-gene statistical methods that predict associations in a GWAS context. We used simulation methods to learn that there is no single, most powerful method. If the properties of the gene model are not known, the most powerful approach is a composite test that uses a recessive-dominant-additive composite model. We also found that regardless of whether the MOI is known or not, there always exists a method that outperforms (in a statistical power context) the Pearson χ^2 test.

Chapter 5: The Influence of Errors Inherent in GWAS in Relation to Single-Gene Models

Chapter 5 is based on a study that was published elsewhere.[28] This chapter describes our investigation of the effects of errors in both genotype and phenotype misclassifications. The central objective is to assess the impact that these errors have on the additional sample size required to achieve a specific power threshold, which for this study is set to 80 percent. Usually, GWAS are conducted assuming that the study measurements are error free. This chapter discusses the assembled evidence challenging that assumption and the examples we used to assess the consequence of those assumptions.

We simulated the effects of genotype errors by intentionally mislabeling the genotype X percent of the time, where X is an exogenously provided model parameter. It was our assumption that genotype errors due to incorrect chip assignments affect both genotypes by incorrectly switching the designation of the disease versus nondisease genotype. We used a similar approach to process phenotype errors that are assumed to affect disease diagnosis. In this process, the simulation data will be recoded to simulate a diagnosis switch from a positive to a negative outcome and vice versa Y percent of the time, where Y is also an exogenously provided user parameter. In general, we assumed that the value of X was less than 1 percent. We posed a much higher value for Y because estimates in the literature report disease misdiagnoses could be as high as 29 percent.[29] These features allowed us to assess the individual or combination of genotype and phenotype error levels and their statistical power consequences. Studies of genotype error have led to a number of investigations in the statistical genetics literature that we review in Chapter 5. Although the accuracy of the genotyping process has improved, errors still occur.

Phenotypic misclassification errors are also a source of bias and can reduce the power of detecting a statistical association between a phenotype and a specific allele.[30,31] To help provide insight into the influence of simultaneous genotype and diagnosis errors affecting the accuracy of the phenotype measure in a GWAS, we ran simulations with synthetically generated data. We focused on assessing how statistical power was affected by the influence of these frequently overlooked sources of errors in GWAS. Our simulations demonstrated that genotype (even at low error rates) and phenotype (diagnosis) errors produce substantial power losses for all MOIs, with significant power losses for recessive MOIs. Because GWAS involving recessive loci have additional power requirements relative to other MOI types, researchers need to address these requirements in developing appropriate sample sizes for their studies.

In summary, this chapter identifies the significant role that epigenetics effects and diagnosis misclassifications can play in designing tests with realistic power levels.

Chapter 6: Conducting GWAS Epistasis Scenarios

Chapter 6 is based on a study that was published previously.[32] This chapter presents the results of our investigations of analyzing epistatic scenarios in GWAS. We used a qualitative association model to assess the statistical models that reliably predict associations between a qualitative phenotype (i.e., a disease diagnosis) and a pair of interacting genes. We employ the concept of relative risk, which is the ratio of the probability of a positive diagnosis given a mutated genotype divided by the probability with no risk present. We used a simulation approach to generate synthetic data corresponding to a variety of possible epistatic models (EMs). Our method took into account the strength of association, disease prevalence in nonrisk populations, and most importantly, the inheritance patterns of the pair of epistatic genes. We analyzed the simulated gene data to assess how these individual factors influenced statistical power in the context of GWAS.

The results indicated that the most powerful statistical methods for predicting associations between phenotypes and genotypes in epistatic scenarios are statistical models that simultaneously test for associations involving both interacting loci. This has significant computational implications. The number of single SNP evaluations is as large as 1,000,000 sets of calculations. The number of SNPs pairs is approximately 5×10^{11} calculations. This result is not

surprising and has been reported by others. An additional significance of our study is that it incorporates new statistical methods as part of the comparison analysis. We also documented the extent to which single-gene models fail to predict associations involving interacting genes with phenotypes constructed to be associated with low risk. Also, each gene MOI affects the ability to identify the association, which further confounds the GWAS methodology.

Chapter 7: Assessing Gene-Environment Interactions in GWAS: Statistical Approaches

Chapter 7 is based on a study that was previously published by RTI Press.[33] Environmental influencing factors on GWAS are described in Chapter 7. Classical statistical tests derived from case-control experiments can be used to determine if two loci associate in a GWAS context. But this model depends on a narrow range of environmental submodel formulations. In this scenario, logistic regression models are versatile approaches because they are able to examine main effects, pairwise interaction assumptions, or both. One early study investigating gene-gene interactions showed that explicitly modeling interactions between loci for GWAS with hundreds of thousands of markers is computationally feasible.[34] In this chapter, we also show that simple methods explicitly considering interactions can actually achieve reasonably high power with realistic sample sizes under different interaction models with some marginal effects, even after adjusting for multiple testing using the Bonferroni correction.

In this chapter, we also focus on low-effect/rare-variant loci with low relative risks of association with disease diagnosis. The overarching goal is to identify which statistical methods best identify genotype-phenotype associations when environmental effects also influence the association. Detecting such associations is particularly difficult for genetic variants with modest impacts on risk. Consequently, our experiments specifically investigated scenarios involving low-risk genetic variants and assessed whether environmental influences with varied levels of risk could be a source of the "missing heritability" observed using single-gene models.[35] Not surprisingly, our investigations demonstrated that the best statistical method (with respect to statistical power) depends on whether there are interactions between the genotype and environmental factors as well as how well the specified statistical model matches the environmental effect associated with the phenotype.

Chapter 8: Polygene Methods in GWAS

Chapter 8 proposes a novel strategy for studying the association between large sets of SNP predictors and groups of correlated phenotypes (i.e., outcomes). These data commonly arise in GWAS, in which associations with a large number of qualitative and quantitative phenotypes are investigated using hundreds of thousands of genetic markers. Our strategy is formulated within the linear–log linear models, a framework suited to the analysis of qualitative trait responses.

Our approach is based on the assumption that single-locus models do not detect all of the markers that are part of the phenotype pathway. In Chapter 6, we showed that for a given locus, single-locus tests are not as effective as two-locus tests. Despite these empirical arguments, which are based on computational considerations, single-gene models remain the core method for detecting associations in a GWAS context. Accordingly, our general strategy is to make an initial pass against all usable SNP autosomes and apply a significance threshold to identify highly significant SNP-phenotype associations, known as stage 1 SNPs. The second step statistically combines the stage 1 SNPs with all original autosome SNPs to identify significant SNP pairs that are phenotype-associated. This step uses a test for significance that is conditional on the stage 1 SNP. We then propose to continue this process for triple SNPs, quadruple SNPs, and so on until the combination of loci produces no new SNP-phenotype associations. This process is analogous to a stepwise regression process, in which networks of SNPs are combined stage by stage until no new SNPs exceed the significance threshold.

Chapter 9: Conclusions

Our final chapter discusses statistical power properties in the GWAS context and offers guidance on selecting "best methods." Our early concerns with the absence of standards led us to develop a synthetic gene database that recorded known outcomes between synthetic phenotypes and genotype networks provides a mechanism that was not possible using real genomics data. The creation of this database enabled an evaluation of different statistical models and methods specifically because the prediction outcomes were known and statistical power profiles could be estimated. We also investigated methods that examine combinations of genes acting in concert either with other genes or environmental factors. Our assessments indicated that often single-gene models will fail to identify markers in many types of gene-gene, gene-environment networks.

We also developed a general polygene test (Chapter 8) that builds up a network of SNPs that link to a single phenotype. A reliable single-gene model is still necessary to identify a starting SNP in the polygene process. However, during this initial pass, the inheritance properties of the genes to which the SNPs belong are predicted with high reliability. We then use the inheritance properties of the SNPs in the step-by-step association process. Knowing the inheritance improves the performance with respect to the statistical power of our polygene process.

Chapter References

1. Johannsen WL. Arvelighedslærens elementer (the elements of heredity). Copenhagen, Denmark: Gyldendalske Boghandel Nordisk Forlag; 1905.
2. Alberts B, Johnson A, Lewis J, et al. Molecular biology of the cell. 4th ed. New York, NY: Garland Science; 2002.
3. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005;308(5720):385-9.
4. Burdett T, Hall P, Hasting E, et al. The NHGRI-EBI Catalog of published genome-wide association studies. 2015 [cited 2015 Nov 2]; Available from: www.ebi.ac.uk/gwas
5. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine. Online Medelian Inheritance in Man (OMIM). 2016 [cited 2016 Feb 11]; Available from: http://www.ncbi.nlm.nih.gov/omim
6. Altmuller J, Palmer LJ, Fischer G, et al. Genomewide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet. 2001;69(5):936-50.
7. Strachan T, Read A. Human molecular genetics. New York, NY: Garland Science; 2010.
8. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996;273(5281):1516-7.
9. Ardini MA, Pan H, Qin Y, et al. Sample and data sharing: observations from a central data repository. Clin Biochem. 2014;47(4-5):252-7.
10. Greely HT. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. Annu Rev Genomics Hum Genet. 2007;8:343-64.

11. Klein C, Lohmann K, Ziegler A. The promise and limitations of genome-wide association studies. JAMA. 2012;308(18):1867-1868.

12. National Human Genome Research Institute. Genome-wide association studies. 2007 [cited 2014 Mar 22]; Available from: http://www.genome.gov/20019523

13. Macarthur D. Bioscience Resource Project critique of modern genomics: a missed opportunity. 2010 [cited 2015 Nov 2]; Available from: http://www.wired.com/2010/12/bioscience-resource-project-critique-of-modern-genomics-a-missed-opportunity/

14. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106(23):9362-7.

15. Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell. 2008;135(2):216-26.

16. Wagner MJ. Rare-variant genome-wide association studies: a new frontier in genetic analysis of complex traits. Pharmacogenomics. 2013;14(4):413-24.

17. Hu X, Daly M. What have we learned from six years of GWAS in autoimmune diseases, and what is next? Curr Opin Immunol. 2012;24(5):571-5.

18. Hirshhorn JN. Genomewide association studies—illuminating biologic pathways. *N* Engl J Med. 2009;360(17):1699-1701.

19. Goldstein DB. Common genetic variation and human traits. *N* Engl J Med. 2009;360(17):1696-1698.

20. Kraft P, Hunter DJ. Genetic risk prediction--are we there yet? *N* Engl J Med. 2009;360(17):1701-1703.

21. Schymick JC, Scholz SW, Fung HC, et al. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol. 2007;6(4):322-8.

22. Cooley P, Clark R, Folsom R, et al. Genetic inheritance and genome wide association statistical test performance. J Proteomics Bioinform. 2010;3(12):321-325.

23. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012;8(12):e1002822.

24. Freidlin B, Zheng G, Li Z, et al. Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered. 2002;53(3):146-52.

25. Sasieni PD. From genotypes to genes: doubling the sample size. Biometrics. 1997;53(4):1253-61.

26. Johnson RC, Nelson GW, Troyer JL, et al. Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics. 2010;11:724.

27. Li Q, Zheng G, Li Z, et al. Efficient approximation of P-value of the maximum of correlated tests, with applications to genome-wide association studies. Ann Hum Genet. 2008;72(Pt 3):397-406.

28. Cooley P, Clark RF, Page G. The influence of errors inherent in genome wide association studies (GWAS) in relation to single gene models. J Proteomics Bioinform. 2011;4:138-144.

29. Kircher T, Nelson J, Burdo H. The autopsy as a measure of accuracy of the death certificate. *N* Engl J Med. 1985;313(20):1263-9.

30. Gordon D, Finch SJ, Nothnagel M, et al. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. Hum Hered. 2002;54(1):22-33.

31. Barendse W. The effect of measurement error of phenotypes on genome wide association studies. BMC Genomics. 2011;12:232.

32. Cooley P, Gaddis N, Folsom R, et al. Conducting genome-wide association studies: epistasis scenarios. J Proteomics Bioinform. 2012;5(10):245-251.

33. Cooley PC, Clark RP, Folsom RE. Statistical methods that identify genotype-phenotype associations in the presence of environmental effects. RTI Press Publication No. RR-0022-1405. Research Triangle Park, NC: RTI Press; 2014.

34. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet. 2005;37(4):413-7.

35. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.

CHAPTER 2

Genome-Wide Association Data: Where Are the Standards?

Philip Chester Cooley

Introduction

In this chapter we assess the predictive strength of a number of classical statistical methods by applying them to a publically available set of amyotrophic lateral sclerosis (ALS) data reported in a paper by Schymick and colleagues (2007).[1] We used these methods in the context of a single locus genome wide association study (GWAS) experiment. The methods are compared and the degree of similarity/dissimilarity between them is empirically measured to determine if a combination of methods is more predictive of phenotype genotype associations than a single method. All of the methods in our assessment are single nucleotide polymorphism (SNP) based.

There are three types of ALS: classic sporadic, familial, and the Mariana Island forms. Classic ALS accounts for 90 to 95 percent of ALS patients in the United States and is called "sporadic" because it cannot be traced to ancestors with the illness.[2] The literature identifies variations and mutations in many Mendelian loci and genes that potentially cause different forms and subtypes of ALS, indicating that many complex and diverse molecular mechanisms are involved in ALS pathogenesis. These genes include SOD1, ALS2, SETX, and VAPB for familial ALS, and VEGF, ANG, HFE, SMN, and PON1 for sporadic ALS. Research has also reported that the inheritance pattern varies with the type of ALS, including autosomal dominant, autosomal recessive, X-linked dominant and maternal through mitochondrial genes.[3-5]

Despite these complexities, ALS researchers have intensified their investigations of sporadic ALS. Researchers are applying the GWAS approach, looking for genes that increase susceptibility to sporadic ALS. By 2008 when the authors' investigations began, seven teams had reported results from ALS-based GWAS,[6-12] but none had highlighted genes already under suspicion,

and each team had reported a different set of ALS markers. Subsequently, eight additional studies[13-20] have been reported in the literature. These results have led to conjectures that using the GWAS method to search for ALS genes may have to accommodate a spectrum of genes, each of which contribute to ALS in some unknown manner. The failure of the different studies to replicate each other's results also suggests that GWAS may be inconsistently implemented, or population admixtures are obscuring the findings.

The first GWAS for ALS found no SNPs significantly associated with the disease.[1] However, published studies that followed implicated several different ALS genetic markers. For example, Dunckley and colleagues[6] identified an SNP located in the FLJ10986 gene. This study consisted of 1,152 patients diagnosed with sporadic ALS and 1,297 controls. The initial discovery was made on analysis done on 386 cases and 547 controls, all of whom were of European descent and older than 65. A residual sample of 766 cases and 750 controls, as well as a subsample of the data identified in the Schymick study, were used to replicate the discovery analysis.[6] A third ALS study reported that the inositol 1, 4, 5-triphosphate receptor 2 (ITPR2) marker was associated with ALS. This study pooled three European populations with 1,337 ALS patients and 1,356 controls.[7] A fourth study (by the same team as the third study) identified an SNP in the dipeptidyl-peptidase 6 (DPP6) gene that was strongly associated with ALS susceptibility.[8] However, the ITPR2 marker was no longer significant. A fifth study examined an Irish population, which was augmented with a Dutch and a US population consisting of 958 ALS cases and 932 controls, and confirmed an association with the DPP6 marker.[21] However, a sixth study sought to confirm the DPP6 marker finding by examining the Irish ALS cohort data and augmenting it with a Polish cohort. Cronin and colleagues,[21] reported that their analysis of the combined cohorts that consisted of 1,267 cases and 1,336 controls was unable to identify any associations including the previously reported DPP6 marker.[21] A seventh study, performed by Chiò et al.,[9] used a two-stage analysis that consisted of 553 cases and 2,338 controls: it identified two new markers, but markers mentioned previously, including DPP6.

Our initial assessment concluded that clear and definitive disease associations from these seven significant ALS studies was lacking and suggested difficulties of using GWAS approaches for complex diseases like ALS. Consequently, we made the decision at that time to pursue simulation-based studies with known outcomes in an effort to explore the possibility of developing standard procedures for conducting GWAS.

Subsequent to our decision to pursue simulation methods to assess GWAS-appropriate methods, nine additional ALS related GWAS were identified in the literature. Blauw et al.[10] investigated the role of copy number variants (CNVs) as a source of genetic variation for 406 ALS cases and 406 controls and were unable to identify a locus associated with sporadic ALS. Another study by Landers et al.[22] implicated the KIFAP3 gene that was associated with increased survival in sporadic ALS. An additional GWAS by van Es et al.[23] used two cohorts to implicate an SNP (rs12608932) located in the UNC13A gene. This gene is known to code proteins that are presynaptic proteins found in central and neuromuscular synapses that regulate the release of neurotransmitters, peptides, and hormones. This same study also showed genome-wide significance for two additional SNPs located in chromosome 9p21. A GWAS by Laaksovirta et al.[24] was able to confirm that genes on chromosome 9p21 and suggested that it could be a major cause of familial ALS in a Finnish population. A further study by Shatunov et al.[25] provided additional evidence that two SNPs in a locus on chromosome 9p21 were associated with ALS. The International Consortium on Amyotrophic Lateral Sclerosis[26] conducted a meta-analysis of ALS samples consisting of 4,243 ALS cases and 5,112 controls from 13 European and US cohorts and provided additional evidence for the loci on UNC13A and chromosome 9p21. However, a study by Daoud et al.[27] could not confirm this result in a relatively small population (285 cases and 285 controls) of French ALS subjects. As a result of increased incidence of ALS in US veterans, Kwee et al.[28] conducted a GWAS of ALS outcome and survival time in a sample of US veterans. They report no SNPs reached genome-wide significance in the discovery phase for either phenotype. A final GWAS by Deng et al.[29] on a Chinese population reported finding two additional susceptibility loci that were not reported in other studies.

Methods

Our study assessed the statistical methods that have appeared in the GWAS literature and included a number of established methods used by the cited studies. We applied each of these methods to the ALS data of Schymick and colleagues[1] and provide a method of comparing their relative performance.

The Schymick et al. (2007) ALS data set

This SNP data was produced by the Laboratory of Neurogenetics of the intramural program of the National Institute on Aging (NIA), National Institutes of Health (NIH). The genotyping was performed using the Illumina

Infinium assay humanhap550. Infinium assays assess haplotype tagging SNPs based upon Phase I+II of the International HapMap Project. The genotype data we used in this study consists of 555,352 SNPs from 276 ALS patients and 271 neurologically normal controls. These data are publicly available to the scientific community.[1]

The Statistical Tests

In a simple GWAS analysis, we rely on individual statistical tests of each typed SNP to identify potential associations. For any of the statistical methods considered for measuring associations, we can represent the sample genotype data or the sample allele data in a contingency array stratified by cases and controls. Under a null hypothesis of no association with the phenotype, we expect no difference between the frequencies across cases and controls. In general, this is the strategy behind all of the statistical methods considered. However, the methods make different operational assumptions that produce very different measures of association.

Testing a person's DNA for more than half a million SNPs will produce many spurious associations. For example, if there were no actual associations, 500,000 independent tests using a $p<.0001$ criterion would be expected to identify 50 candidate genes. Statistical methods can correct for this, but they can also obscure real associations and produce both false positive and false negative associations. For this study, we are using the data generated by the Illumina 550K chip, which has an SNP about every 5 kb. Also, loci near each other might not be mutually independent. The general strategy for protecting against type I errors is by setting a stringent statistical significance threshold.

Each locus is screened to establish if there is sufficient information to apply the statistical procedures. The method we used is based on the procedure defined by Zeggini and colleagues.[30] SNPs are considered eligible if the minor allele frequency (MAF) exceeds 1 percent in both cases and controls. These restrictions help protect against computational difficulties caused by data sparseness in our statistical calculations. Another purpose of any eligibility test is to determine if there is sufficient representation of all genotypes to perform an accurate statistical test.

We investigated a number of statistical tests to use in our assessment including the following.

The Case-Control Genotype Method, Based on the Pearson χ^2 Test. This classic test is used in many epidemiological studies. The procedure constructs a 2×3 genotype table (case control by the three genotypes) that uses a Pearson χ^2 test

with 2 degrees of freedom to test the hypothesis that the cases and controls are from the same distribution. Under the null hypothesis of no association with disease, we expect the relative genotype frequencies to be the same in cases and controls. These types of methods are commonly used in the GWAS context. A case-control study uses the odds ratio to estimate the relative risk and assumes that the disease under study has a low incidence. When the risk ratio is the parameter of interest, the assumption of rarity is needed for the odds ratio to be a consistent estimator.[31,32] This method was used in the original Schymick study.

Normal Approximation to Fisher's Exact Test—Dominant and Recessive Models. The null hypothesis behind Fisher's test is that the rows (phenotype) and columns (genotypes) are unrelated. The test calculates an exact probability value for the relationship between three dichotomous variables, as found in a 2×3 table. When N (the number of subjects) is large, the exact form of the Fisher test is difficult to calculate. Therefore, a normal approximation is used. Because the test estimates the probability of a given genotype using the marginal values and assumes the probability is from a normal curve, an autosomal recessive and dominant version of the test is easily implemented. We used both forms in our assessment, listed as Fis-D and Fis-R in Table 2.1.

Logistic Regression Linear and Categorical Model Tests. The logistic regression test (Log-A) assumes an additive mode of phenotype inheritance and regresses case control outcomes using the number of minor alleles as the dependent variable. In the Log-A test, the null hypothesis $\beta 1 = 0$ is used to test if the number of alleles associates with the phenotype variable.[33]

Table 2.1. Correlation values for eight statistical tests based on unadjusted *p*-values

Test Name	Fis-D	Pea	Tr-A	All	Fis-R	Tr-D	Log-A	Tr-R
Fis-D	1.0	-.0472	-.0544	-.0535	0.2215	-.0947	-.0551	-.0041
Pea	-.0472	1.0	0.4861	0.4832	-.0651	0.4818	0.4874	0.4793
Tr-A	-.0544	0.4861	1.0	0.9843	-.0105	0.5294	0.9973	0.0821
All	-.0535	0.4832	0.9843	1.0	-.0104	0.5269	0.9834	0.0811
Fis-R	0.2215	-.0651	-.0105	-.0104	1.0	-.0026	-.0107	-.1677
Tr-D	-.0947	0.4818	0.5294	0.5269	-.0026	1.0	0.5299	0.0020
Log-A	-.0551	0.4874	0.9973	0.9834	-.0107	0.5299	1.0	0.0822
Tr-R	-.0041	0.4793	0.0821	0.0811	-.1677	0.0020	0.0822	1.0

Fis-D = Fisher dominant test; Pea = Pearson χ^2 test; Tr-A = trend additive test; All = allelic test; Fis-R = Fisher recessive test; Tr-D = trend dominant test; Log-A = logistic linear test; Tr-R = trend recessive.

Cochran-Armitage Trend Tests: Recessive, Additive, and Dominant Models. The classical Cochran-Armitage test assumes an additive mode of inheritance (MOI) and is typically used in categorical data analysis when some categories are ordered. The test is sensitive to the linearity between phenotype and genotype variables and detects trends that would not be noticed by other tests.[34] It also uses weights applied to each genotype variable to generalize the method. For example, a weight of 0.0, X, or 1.0 (where $0 \geq X \leq 1$) can be used to assume different MOI assumptions: recessive (X = 0), additive (X = .5) or dominant (X = 1) locus. The method is discussed by Zheng and Gastwirth.[35] We use all three test assumptions in our assessment and identify them as Tr-R, Tr-A, and Tr-D, respectively, in Table 2.1.

Allele Test. This is a commonly used test for association in a 2 × 2 contingency table, in which cases and controls are classified as carriers of the minor, risk-carrying allele. This is a 1 degree of freedom (1*df*) test that assumes dominance on a single allele. This test constructs a 2 × 2 allele table (case control by the two alleles) and uses a Pearson χ^2 (CHI2) test (1*df*) to test the hypothesis that the cases and controls are from the same distribution.[32]

Results

A fundamental question this study seeks to address is "which method should investigators use to assess candidate associations?" No one has yet answered this question. One obvious partial answer is that it depends on the gene behavior, as well as a number of other biological factors yet to be determined. Consequently, if we assume an additive gene model, then the likelihood of establishing the association between genotype and phenotype will be nearly the same whether or not one uses any of the three additive-based tests.

Test Performance

In an attempt to address the question above, we created a 2 × 3 table of counts of case-control subjects by genotype counts for each locus. We then added the marginal values and provided the necessary information to apply the eligibility criteria to eliminate loci with low minor genotype representation. The data provides information on 555,352 loci, of which 538,234 are autosome loci. Applying a standard quality control method described by Zeggini et al.[30] to the autosome loci identifies 55,304 loci (10 percent) that were screened ineligible and not considered in the analysis due to low minor genotype representation. Table 2.1 presents the 8 × 8 matrix of Pearson product-moment correlation coefficients for all possible pairs of the eight statistical tests. Note

that the statistical tests differ in the MOI assumption and therefore should provide different results. However, Table 2.1 indicates that although some of the MOI-based tests (i.e., the additive tests) produce consistent results, the dominant based tests (Tr-D and Fis-D: correlation coefficient = -.0947), and the recessive based tests (Tr-R and Fis-R correlation coefficient = -.1677) are not correlated. The Pearson test is MOI agnostic and would be expected to overlap with the other tests, and although they both capture 48 percent of the correlation space represented by many of the other tests, they have only a small negative correlation with both of the Fisher tests. Thus, Table 2.1 illustrates that all of the tests historically used in GWAS incorporate different assumptions and consequently have differences in the way they measure association. Based on results presented here, we assert that if the MOI characteristics of the associated loci are unknown, then researchers should consider multiple tests treated in a nonhierarchal manner.

In addition, Table 2.1 suggests a correlation between the three tests (Tr-A, Log-A, and All) that assume additive MOI properties. This suggests that no additional predictive power for measuring associations is derived from using more than one of these three tests; that weak correlation between the two tests (Tr-R and Fis-R) that assume recessive MOI properties and weak negative correlation between the Fisher dominant MOI test (Fis-D) and all of the other tests except the Fisher recessive MOI test (Fis-R), which implies that these tests measure a dimension that is different from all of the other tests.

Marker Assessment

This section compares our results directly to those of Schymick and colleagues[1] and indirectly to other results reported in the literature.

Schymick et al. (2007). The Schymick and colleagues study[1] reports using six association tests: the genotypic test, two versions of the trend test (the dominant and additive tests), a recessive model test, an allele-based test, and a three-marker haplotype-association test. We included their five-loci-based tests in the tests we ran.

We applied the Pearson test, the recessive version of the trend test, and the additive version of the logistic regression test to the same data. A summary of these results is presented in Table 2.2 below. It compares the top 34 SNPs reported by Schymick and colleagues[1] with our results. We found that all 34 were positive at the e^{-4} level of significance based on the Pearson test but that none were significant at the 10^{-7} level. Also, 13 of the SNPs were positive at the

Table 2.2. Comparison between the results of our study and the Schymick et al. study

SNP ID	Chrom.[$]	Location	Gene	Pea	Tr-R	Log-A
rs4363506	10q26.13	129164493	Intergenic	<.000001	<.005	<.000001
rs16984239	2p24	18097927	Intergenic	<.00001	X	<.00001
rs12680546	8q24.2	136940921	Intergenic	<.0001	X	<.001
rs6013382	20q13.2	50136040	ZFP64	<.00001	X	<.01
rs2782931	9q31.3	113890011	SUSD1	<.00001	<.0001	X
rs11099864	4q31.3	154112804	KIAA1727	<.00001	<.0005	X
rs332389	3p14.1	$66493904	SLC25A26	<.0001	<.01	X
rs4964213	12q23.3	106274907	BTBD11	<.0001	<.0001	<.0001
rs10765118	10q26.13	129175173	Intergenic	<0001	<.00001	<.0001
rs3733242	4q21.1	77894529	SHROOM3	<.0001	<.00001	<.0001
rs1037666	1q43	238425108	FMN2	<.0001	X	<.001
rs1436918	15q14	32724213	LOC390569	<.0001	<.0001	X
rs4552942	8q24.2	136943505	Intergenic	<.0001	X	<.001
rs852801	1p32.2	58094497	DAB1	<.0001	<.00001	<.001
rs852802	1p32.2	58096531	DAB1	<.0001	<.00001	<.001
rs7250467	19q12	33261241	LOC727771	<.0001	<.0001	<.0001
rs10830099	10q26	129174355	Intergenic	<.0001	<.0001	< 0001
rs10459680	15q26	91482474	Intergenic	<.0001	X	<.001
rs1752784	9q22.32	96217647	HIATL1	<.0001	<.00001	<.01
rs1202824	1p32.2	58121593	DAB1	<.0001	X	<.01
rs5014235	5q14.1	77245417	Intergenic	<.0001	<.01	<.0001
rs7201419	16q23.3	81887480	CDH13	<.0001	<.0001	<.01
rs11933187	4q31.3	175446507	KIAA1717	<.0001	<.0001	<.001
rs10773543	12q24.32	127489679	TMEM132C	<.0001	X	<.001
rs7976059	12q13	50537539	Intergenic	<.0001	<.0001	<.001
rs9608416	22q12.1	24441018	ADRBK2	<.0001	<.01	X
rs2220999	12q12	40422035	Intergenic	<.0001	<.01	<.0001
rs12632457	3p24	27995556	Intergenic	<.0001	<.01	<.0001
rs2272519	2p24	18575231	Intergenic	<.0001	<.05	X
rs2289599	5q14.1	77243905	Intergenic	<.0001	<.05	<.0001
rs4478530	8p12	31517686	Intergenic	<.0001	<.05	X
rs130110	22q13.32	47470399	FAM19A5	<.0001	X	<.0001
rs9510982	13q12	23451861	Intergenic	<.0001	<.0001	X
rs2767584	6p21	156964439	Intergenic	<.0001	X	X
Total < e^{-4}				**34**	**13**	**12**

Fis-D = Fisher dominant test; Pea = Pearson χ^2 test; Tr-A = trend additive test; All = allelic test; Fis-R = Fisher recessive test; Tr-D = trend dominant test; Log-A = logistic linear test; Tr-R = trend recessive.

[$]Chromosome location of single nucleotide polymorphism (SNP).

Some data from Schymick et al.[1]

10^{-4} criteria according to the Trend-R test and 12 were positive according to the Logistic-A test.

Thus, 12 of 34 association tests at the e^{-4} level overlap between the MOI-agnostic Pearson test and both the recessive Tr-R test and the Log-A test. We can therefore replicate the Schymick et al. results but only if we use the Pearson test, which is commonly used for GWAS. The Trend-A test is equally popular. However, if another MOI-specific test had been used instead, we could not have replicated the results.

Other ALS GWAS. The year we first investigated the Schymick et al.[1] ALS GWAS study, there were seven other GWAS also published on ALS. These studies were Dunckley et al.[6]; van Es et al.[7,8]; Blauw et al.[10]; Cronin et al.[11,12]; and Chiò et al.[9] The focus of the Blauw et al.[10] study was unique. It investigated copy number variations, and the reported results were negative. We examined the top markers reported in the remaining six studies and summarize that information in Table 2.3. The DPP6 marker first identified by van Es et al.[8] is worth mentioning because it was also reported in one of the Cronin et al. studies.[11]

Table 2.3. Association test results of SNPs identified as significant in other studies

Study	SNP ID	Chr	Gene	*p*-value reported	*p*-value this study
Dunckley[6]	rs6700125	1	FLJ10986	$1.8\ 10^{-5}$	$3.1\ 10^{-3}$
Dunckley[6]	rs6690993	1	FLJ10986	$2.0\ 10^{-4}$	NA
van Es[7]	rs2306677	12	ITPR2	$7.0\ 10^{-4}$	X
van Es,[8] Cronin[11]	rs10260404	7	DPP6	$5.04\ 10^{-8}$	$7.4\ 10^{-4}$
Chio[9]	rs2708909	7	SUNC1	$6.98\ 10^{-7}$	$4.5\ 10^{-3}$
Chio[9]	rs2708851	7	Intergenic	$1.16\ 10^{-6}$	$9.5\ 10^{-3}$

NA = SNP not included on chip; X = not significant at any level; SNP = single nucleotide polymorphism.

The results summarized in Table 2.3 identify a *p*-value < 10^{-3} for one of the markers reported in other studies; however, all of the five others are less significant. At this threshold, we are unable to replicate any other reported results and therefore must hypothesize that the collective ALS studies have identified no predisposing biological markers. We also performed a linkage disequilibrium (LD) analysis of all SNPs in Table 2.3 to identify any other SNP close to the index SNP in column 2 (i.e., < 200kb of the index SNP) with high LD. We were unable to identify any.

Finally, we examined the PON1 gene that was reported to be associated with sporadic ALS in genetic studies by more than one study. [36,37] All tested nonsignificant.

Conclusions

We used publicly available data that contained 276 cases and 271 controls. This sample is very underpowered, with many SNPs containing few disease alleles. The SNP coverage may also have been insufficiently dense. Although the SNPs on the chip are tag SNPs that were selected to represent all the genes in a comprehensive manner, the selection may not have included all the potential markers for a specific disease. Other explanations are that the sample of individuals was too heterogeneous, resulting in the study populations with distinct disease penetrance traits; the strength on the association was too weak to be detected; errors in the data perturbed the measurements; or the phenotype definition differed across studies. Novel statistical methods cannot overcome the problems inherent in poor quality data containing too few subjects, or data that uses a poorly defined phenotype (i.e., an ALS diagnosis).

We found that different statistical tests varied significantly in estimating the test association measurement, implying that GWAS results depend on the chosen statistical method. We also find that there is no compelling standard that establishes which statistical methods investigators should use in the context of GWAS with unknown MOI properties. Although GWAS have used and reported findings for a number of different tests, most tests have unique properties and consequently can prescribe a different candidate set of phenotype SNP associations. In general, the *p*-value threshold is Bonferroni corrected to a very small value, which encourages high type II error rates.[38]

GWAS hold substantial promise, but for many phenotypes, the path forward is complicated for many reasons not yet understood. The replicability of results in some but not all studies suggests that researchers should present any and all conclusions with caution. Identifying associations solely on the basis of extreme *p*-values is likely to be misleading because an extreme *p*-value alone does not identify the underlying biological mechanism that produces the association. Furthermore, the rate of missing genotype measures suggests that the genotype data contains an unknown number of errors (e.g., in the case of ALS, obtaining an accurate diagnosis is notoriously difficult). Accurate error rates are important components for assessing statistical power properties, and their absence will lead to underpowered GWAS.

Chapter References

1. Schymick JC, Scholz SW, Fung HC, et al. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol. 2007;6(4):322-8.
2. MedicineNet.com. Amyotrophic lateral sclerosis (ALS or "Lou Gehrig's Disease"). 2015 Nov 23 [cited 2015 Nov 23]; Available from: http://www.medicinenet.com/amyotrophic_lateral_sclerosis/article.htm
3. Orrell RW. Understanding the causes of amyotrophic lateral sclerosis. *N* Engl J Med. 2007;357(8):822-3.
4. Hardiman O, Greenway M. The complex genetics of amyotrophic lateral sclerosis. Lancet Neurol. 2007;6(4):291-2.
5. Pasinelli P, Brown RH. Molecular biology of amyotrophic lateral sclerosis: insights from genetics. Nat Rev Neurosci. 2006;7(9):710-23.
6. Dunckley T, Huentelman MJ, Craig DW, et al. Whole-genome analysis of sporadic amyotrophic lateral sclerosis. N Engl J Med. 2007;357(8):775-88.
7. van Es MA, Van Vught PW, Blauw HM, et al. ITPR2 as a susceptibility gene in sporadic amyotrophic lateral sclerosis: a genome-wide association study. Lancet Neurol. 2007;6(10):869-77.
8. van Es MA, van Vught PW, Blauw HM, et al. Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. Nat Genet. 2008;40(1):29-31.
9. Chio A, Schymick JC, Restagno G, et al. A two-stage genome-wide association study of sporadic amyotrophic lateral sclerosis. Hum Mol Genet. 2009;18(8):1524-32.
10. Blauw HM, Veldink JH, van Es MA, et al. Copy-number variation in sporadic amyotrophic lateral sclerosis: a genome-wide screen. Lancet Neurol. 2008;7(4):319-26.
11. Cronin S, Berger S, Ding J, et al. A genome-wide association study of sporadic ALS in a homogenous Irish population. Hum Mol Genet. 2008;17(5):768-74.
12. Cronin S, Greenway MJ, Prehn JH, et al. Paraoxonase promoter and intronic variants modify risk of sporadic amyotrophic lateral sclerosis. J Neurol Neurosurg Psychiatry. 2007;78(9):984-6.

13. Ahmeti KB, Ajroud-Driss S, Al-Chalabi A, et al. Age of onset of amyotrophic lateral sclerosis is modulated by a locus on 1p34.1. Neurobiol Aging. 2013;34(1):357 e7-19.

14. Deng M, Wei L, Zuo X, et al. Genome-wide association analyses in Han Chinese identify two new susceptibility loci for amyotrophic lateral sclerosis. Nat Genet. 2013;45(6):697-700.

15. Kwee LC, Liu Y, Haynes C, et al. A high-density genome-wide association screen of sporadic ALS in US veterans. PLoS One. 2012;7(3):e32768.

16. Laaksovirta H, Peuralinna T, Schymick JC, et al. Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study. Lancet Neurol. 2010;9(10):978-85.

17. Landers JE, Melki J, Meininger V, et al. Reduced expression of the Kinesin-Associated Protein 3 (KIFAP3) gene increases survival in sporadic amyotrophic lateral sclerosis. Proc Natl Acad Sci U S A. 2009;106(22):9004-9.

18. Shatunov A, Mok K, Newhouse S, et al. Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: a genome-wide association study. Lancet Neurol. 2010;9(10):986-94.

19. Sha Q, Zhang Z, Schymick JC, et al. Genome-wide association reveals three SNPs associated with sporadic amyotrophic lateral sclerosis through a two-locus analysis. BMC Med Genet. 2009;10:86.

20. van Es MA, Veldink JH, Saris CG, et al. Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. Nat Genet. 2009;41(10):1083-7.

21. Cronin S, Tomik B, Bradley DG, et al. Screening for replication of genome-wide SNP associations in sporadic ALS. Eur J Hum Genet. 2009;17(2):213-8.

22. Landers JE, Melki J, Meininger V, et al. Reduced expression of the Kinesin-Associated Protein 3 (KIFAP3) gene increases survival in sporadic amyotrophic lateral sclerosis. Proc Natl Acad Sci U S A. 2009;106(22):9004-9.

23. van Es MA, Veldink JH, Saris CG, et al. Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. Nat Genet. 2009;41(10):1083-7.

24. Laaksovirta H, Peuralinna T, Schymick JC, et al. Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study. Lancet Neurol. 2010;9(10):978-85.

25. Shatunov A, Mok K, Newhouse S, et al. Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: a genome-wide association study. Lancet Neurol. 2010;9(10):986-94.

26. Ahmeti KB, Ajroud-Driss S, Al-Chalabi A, et al. Age of onset of amyotrophic lateral sclerosis is modulated by a locus on 1p34.1. Neurobiol Aging. 2013;34(1):357 e7-19.

27. Daoud H, Belzil V, Desjarlais A, et al. Analysis of the UNC13A gene as a risk factor for sporadic amyotrophic lateral sclerosis. Arch Neurol. 2010;67(4):516-7.

28. Kwee LC, Liu Y, Haynes C, et al. A high-density genome-wide association screen of sporadic ALS in US veterans. PLoS One. 2012;7(3):e32768.

29. Deng M, Wei L, Zuo X, et al. Genome-wide association analyses in Han Chinese identify two new susceptibility loci for amyotrophic lateral sclerosis. Nat Genet. 2013;45(6):697-700.

30. Zeggini E, Weedon MN, Lindgren CM, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science. 2007;316(5829):1336-41.

31. Szklo M, Nieto FJ. Epidemiology: beyond the basics. Boston, MA: Jones and Bartlett; 2004.

32. Balding DJ, Bishop M, Cannings C, editors. Handbook of statistical genetics. 3rd ed. Chichester, United Kingdom: John Wiley & Sons; 2007.

33. Garson GD. Logistic regression, in Statnotes: Topics in multivariate analysis. 2008. Available from: http://www2.chass.ncsu.edu/garson/pa765/statnote.htm

34. Agresti A. Categorical data analysis. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2002.

35. Zheng G, Gastwirth JL. On estimation of the variance in Cochran-Armitage trend tests for genetic association using case-control studies. Stat Med. 2006;25(18):3150-9.

36. Mackenzie IR, Bigio EH, Ince PG, et al. Pathological TDP-43 distinguishes sporadic amyotrophic lateral sclerosis from amyotrophic lateral sclerosis with SOD1 mutations. Ann Neurol. 2007;61(5):427-34.

37. Slowik A, Tomik B, Wolkow PP, et al. Paraoxonase gene polymorphisms and sporadic ALS. Neurology. 2006;67(5):766-70.

38. Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006;7(10):781-91.

CHAPTER 3

Creating the Synthetic Gene Data

Philip Chester Cooley

Overview

We used simulation methods to compensate for the absence of both methodological standards and a process for evaluating the statistical methods used in predicting associations between genes and phenotypes. This required creating a synthetic gene database of simulated data linked to known association outcomes constituting a "truth set." Using these data, we analyzed the simulated data using candidate statistical methods for the purpose of assessing each method's predictive properties in the context of an experimental setting that we create.

Our method for generating the synthetic marker data is based on Mendelian concepts of inheritance and epidemiological concepts of relative risk (RR), which is the ratio of the probability of an event occurring in an exposed group to the probability of the event occurring in a comparison, nonexposed group. For example, nonsmokers who inhale secondhand smoke may be more likely to develop lung cancer than nonsmokers who have not been exposed. Individuals also have genetic inheritance elements that include autosomal dominant and autosomal recessive patterns conforming to single-gene inheritance effects. We also incorporate additive and multiplicative inheritance patterns to represent the actions of multifactorial inheritance processes.

We used a study by Iles to represent the contrast between a formal disease diagnosis that stems from genetic causes and the concept of disease penetrance.[1] Penetrance in genetics is the proportion of individuals carrying a particular variation of a gene (allele or genotype) that also express an associated trait. We designate **a** as the risk allele, and **A** as the allele without risk. Generating the synthetic gene data set was facilitated by defining the relationships between penetrance and relative risk for different MOI categories.

This chapter describes a generic process to generate genotype-phenotype data that we use in Chapters 4 through 8. All of the chapters use a version of the data generation process that is derived from this generic process, but there can be differences in detail that depend of the technical content of the chapter.

The Data Generation Process

Specifically the steps were:

1. Preload the details that define the factor combinations for each MOI category. The factors are specified in Table 3.1.

Table 3.1. Factors that define a synthetic gene data file

Factor	Symbol	Number of factors
Sample size	*N*	NCC
Penetrance	P	NP
Phenotype error rate	Perr	NY
Genotype error rate	Gerr	NX
Relative genetic risk	Φ	NGR
Relative environmental risk	Π	NER

2. Draw a genotype distribution at random from the master set of genotype distributions obtained from real distribution data (i.e., the study by Schymick et al.[2]). At this stage, Chan et al. recommends that a minor allele frequency (MAF) threshold not be applied.[3] They argue that filtering MAFs out of the process because of low frequencies or to maintain Hardy–Weinberg equilibrium (HWE) deviation has little effect on the overall false positive rate and in some cases, filtering MAF only serves to exclude SNPs. This step effectively selects a specific genotype distribution (at random) from the master distribution.
3. Use Table 3.2 to assign a case (1) or a control (0) based on the selected genetic relative risk (Φ), penetrance (P) and MOI category. This step converts the Φ ratio value into the probability that the case occurs for the MOI gene model of interest. This process is represented by the following logic that was derived from Iles[1]:

Major Homozygote (AA). Assume that the **AA** genotype is selected. The probability of a case given this selection is equal to the disease penetrance P, or $\mathbf{\Psi_{AA}} = P$.

Minor Homozygote (aa): Liability Increasing Allele. Assume the **aa** genotype is selected. The genetic relative risk ($\mathbf{\Psi_{aa}}$) can be expressed as a ratio of two

probabilities: the probability of a case for a minor homozygote divided by the probability of a case for a major homozygote, or

$$\mathbf{\Psi_{aa}} = \text{Prob(case/}\mathbf{aa}\text{)} / \text{Prob(case/}\mathbf{AA}\text{)} = \text{x/P.} \quad (3.1)$$

From (3.1) the probability of a case given the minor genotype = x =

$$\Psi_{\mathbf{aa}} \times \text{P}, \quad (3.2)$$

where $\Psi_{\mathbf{aa}}$ = one of the assigned risk factors and P is one of the assigned penetrance factors.

Heterozygote (aA). Assume the **aA** genotype is selected. By the same argument, the phenotype risk given a heterozygote is:

$$\mathbf{\Psi_{aA}} = \text{Prob(case/}\mathbf{aA}\text{)} / \text{Prob(case/}\mathbf{AA}\text{)} = \text{y/P.} \quad (3.3)$$

By the same argument, the risk of a case given the heterozygote is

$$y = \mathbf{\Psi_{aA}} \times \text{P}, \quad (3.4)$$

where $\mathbf{\Psi_{aA}}$ = one of the assigned risk factors and P is one of the assigned penetrance factors.

Using the estimate of x and y, assign a case or control at random using the four different MOI models in conjunction with equations (3.2) and (3.4) and Table 3.2. We assigned cases in proportion to x (y) and controls in proportion to 1- x (1-y) for the minor homozygote (heterozygote) genotypes respectively. For the MOI models that assume an elevated risk from the minor and the hetero genotypes, we would expect a higher proportion of cases to be more easily identified via the statistical procedures. The specification of risk depends on specific and unknown disease mechanisms. A relative risk of 1.7 is considered strong and is associated with positive replication,[4] and a risk of 1.3 is considered by Ziegler et al.[5] to be a realistic assumption for complex diseases. In summary, individuals are either assigned as cases or controls according to the probabilities given in Table 3.2.

4. Systematically select subjects. If the subject is a case (control), change its phenotype designation to a control (case) at a rate determined by Perr.
5. Systematically select subjects. If the genotype is a disease (nondisease) allele change the allele to a nondisease allele (disease) allele at a rate determined by Gerr.

6. Continue with the previously described process until *n1* cases and *n2* controls ($N = n1 + n2$) are generated (note that *n1* and *n2*, are not required to be equal).
7. Apply a set of statistical methods to predict associations and record the results.
8. Generate NR (typically NR = 1,000) replicate experiments for each factor combination.
9. Analyze the data.

Table 3.2. Relative risk assumptions, by mode of inheritance

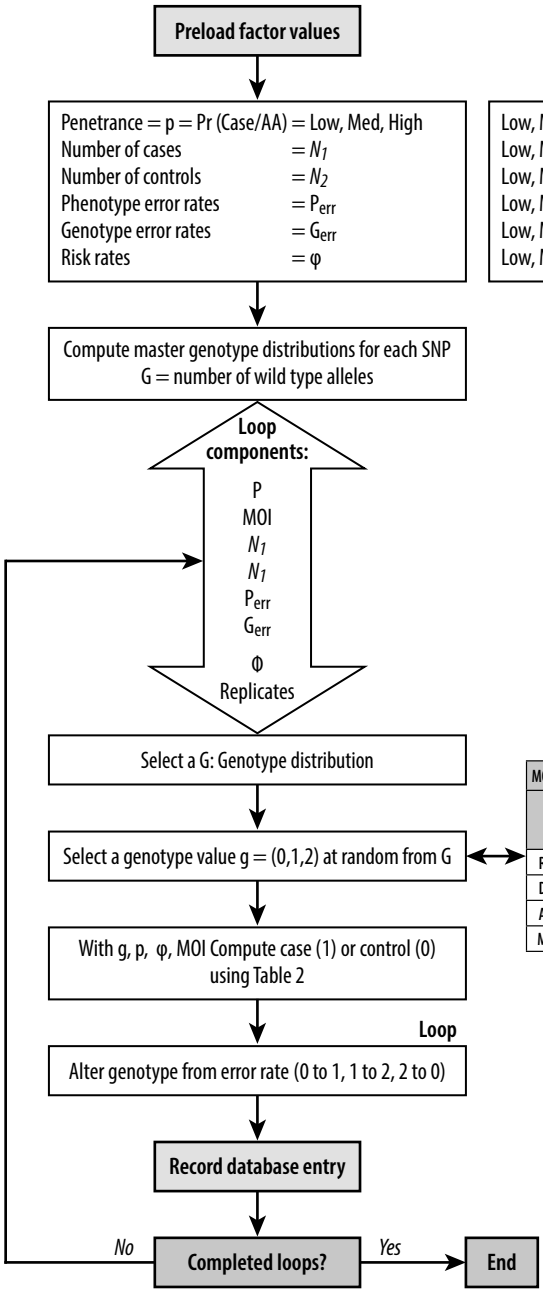
	Major homozygote	Minor homozygote	Heterozygote
Inheritance model	Ψ_{AA}	$\Psi_{aa} = \frac{\Pr(case/\mathbf{aa})}{\Pr(case/\mathbf{AA})}$	$\Psi_{Aa} = \frac{\Pr(case/\mathbf{aA})}{\Pr(case/\mathbf{AA})}$
Recessive	1	Φ	1
Dominant	1	Φ	Φ
Additive	1	2 × Φ-1	Φ
Multiplicative	1	Φ × Φ	Φ

Source: Iles.[1]

- $\mathbf{\Psi_{aa}}$ is the relative risk of homozygous minor to homozygous major.
- $\mathbf{\Psi_{aA}}$ is the relative risk of heterozygote to homozygous major.

Figure 3.1 presents a flow description of the data generation process.

Figure 3.1. Schema of data-generation process



MOI	G = 0 (AA)	G = 2 (aa)	G = 1 (Aa)
	Ψ_{AA}	$\Psi_{aa} = \frac{Pr(case/aa)}{Pr(case/AA)}$	$\Psi_{Aa} = \frac{Pr(case/aA)}{Pr(case/AA)}$
R	1	Φ	1
D	1	Φ	Φ
A	1	2 × Φ-1	Φ
M	1	Φ × Φ	Φ

AA = major genotype; **Aa** = heterozygote genotype; **aa** = minor genotype; MOI = mode of inheritance.

Computational Requirements

Note that the number of unique entity combinations being simulated (NS) as described by the data generation process (see Table 3.1) is:

$$NS = NP \times NX \times NY \times NGR \times NER. \tag{3.5}$$

Each of the NS combinations of traits consists of *N* cases + controls, and each unique combination is replicated NR times. Thus, if the each trait of interest had three risk levels (high, medium, and low) and if NR = 1,000, the number of unique experimental combinations would equal $3^5 = 243$, the number of replicate simulations = 243,000, and the number of entries in the database = $2.43 \times 10^5 \times N$. Depending on the value of *N*, the number of hours of computations required by this process can vary from 2 to 80, and the database entries from 2.43×10^8 ($N = 1{,}000$) to 2.43×10^{11} ($N = 100{,}000$). Note the space requirement on the later assumption is approximately 10 terabytes.

Chapter References

1. Iles MM. Effect of mode of inheritance when calculating the power of a transmission/disequilibrium test study. Hum Hered. 2002;53(3):153-7.
2. Schymick JC, Scholz SW, Fung HC, et al. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol. 2007;6(4):322-8.
3. Chan EK, Hawken R, Reverter A. The combined effect of SNP-marker and phenotype attributes in genome-wide association studies. Anim Genet. 2009;40(2):149-56.
4. Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007;445(7130):881-5.
5. Ziegler A, Konig IR, Thompson JR. Biostatistical aspects of genome-wide association studies. Biom J. 2008;50(1):8-28.

CHAPTER 4

Genetic Inheritance and Genome-Wide Association Statistical Test Performance Using Simulated Data

Philip Chester Cooley, Robert F. Clark,
Ralph E. Folsom, and Grier Page

Overview

Choosing a particular statistical method for a study significantly affects the power profiles of genome-wide association study (GWAS) predictions. Previous simulation studies of a single synthetic phenotype marker determined that the gene model or mode of inheritance (MOI) was a major influence on power. In this chapter, we compare the power profiles of GWAS statistical methods, ones that combine MOI specific methods into multiple test scenarios, against individual methods that may or may not assume an MOI gene model consistent with the marker that predicts the association. Combining recessive, additive, and dominant individual tests are combined and used with either the Bonferroni correction method or the MAX test[2] with respect to single-test GWAS-based methods. If the gene model behind the associated phenotype is not known, a multiple test procedure could have significant advantages compared to single test procedures.

We found that the best statistical method for a study depends on the MOI gene model associated with the phenotype (diagnosis) in question. Our results also indicate that a common assumption that the MOI of the locus associated with the diagnosis is additive will have adverse prediction consequences if the assumption is incorrect.

Chapter 4 is based on a study that was published in the *Journal of Proteomics & Bioinformatics*.[1]
 Chapter 3 of the current publication describes the generation of the synthetic gene database. This section was removed from Chapter 4. The analysis of the gene data has not changed.

Overall, our results indicate that researchers should consider a statistical methodology that combines the results of individual MOI-based core tests for conducting the initial screen in a GWAS. The core tests can be combined into a single operational test in a number of ways. We identify two: the Bonferroni procedure and the MAX procedure, each of which produce very similar statistical power profiles.

Introduction

In this chapter, we examine the statistical methods used to perform GWAS. GWAS usually apply univariate statistical tests to each gene marker or single nucleotide polymorphism (SNP) as an initial step. This SNP based test is statistically straightforward and the testing is done with standard methods (e.g., χ^2 tests, regression) that have been studied outside of the GWAS context. A paper by Kuo & Feingold[3] described the most commonly used methods, and the authors note the use of a compound procedure that combines two or more statistical tests.

The literature also contains a number of papers that compare statistical power among subsets of these methods.[4,5] However, the question of which method is best suited to univariate scanning in a GWAS remains an open issue. The choice of method depends on the match between the true genetic model underpinning the association and the type of model assumed by the method.

To investigate further, we used a multiple test procedure that combined the most promising of the methods identified in the literature and applied them to a set of synthetic marker data with known properties (the Schymick et al. data set introduced in Chapter 2).[6] Our goal was to identify marker properties that could be linked to optimal methods (with reference to statistical power) for predicting associations in GWAS. We know from prior studies that the statistical procedure a researcher chooses influences GWAS prediction accuracy and that there are specific properties of the underlying markers that determine the optimization of the procedure choice.[3]

We also included the important properties that influence the association prediction accuracy into our synthetic marker data via a Monte Carlo simulation process, and we link the properties to the influencing marker to study their individual and collective contributions to association prediction. A synthetic marker data set allows us to assess the performance of different statistical methods in a GWAS context. We applied a number of statistical methods to the simulated data and used their statistical power profiles to

evaluate the performance of the methods. We also quantified the relationships between locus traits and prediction accuracy.

This chapter identifies a number of these properties and quantifies the loss in power if studies use nonoptimal methods. Similar results have been reported in earlier studies.[4,5] Both studies reinforce the view that the major influence on prediction accuracy is the gene model of the locus associated with the diagnosis.

We were particularly interested in assessing the consequences of applying a statistical method that assumes an inherent additive mode of inheritance (MOI) property to nonadditive SNP data. Our motivation for this was twofold. First, the additive MOI model is commonly employed in GWAS; and second, the answer to the question: "what statistical methods should be used to conduct GWAS?" does not have a definitive answer. The best method typically depends on what MOI gene model has been associated with the associated diagnosis.

Our results show the major factors that influence association predictions. They also indicate that a strategy based on predicting associations using multiple statistical methods can be more accurate; much more accurate if the governing marker is recessive, than those that assume a single, additive mode. The multiple test procedure we developed and propose here combines recessive, additive and dominant MOI-optimal statistical methods, all of which are derived from the well-known Cochran-Armitage (CA) test. We also examined different procedures for combining the tests.

Methods

We examined the accuracy of association detection by generating synthetic data with properties that are known to influence statistical power. We used a Monte Carlo process to generate the data from a set of random variables described in Chapter 3. The main purpose of the synthetic data from Schymick et al.[6] is to act as a "truth set" to assess the performance of commonly used statistical methods used in a GWAS context. Please see Chapter 3 for the description of how the data were generated.

The simulated data set that was generated had the following characteristics:

- The proportion of cases (controls) that are major homozygotes = 50.3 (63.0) percent.
- The proportion of cases (controls) that are heterozygotes = 39.2 (31.3) percent.

- The proportion of cases (controls) that are minor homozygotes = 10.5 (5.7) percent.
- With MOI distribution:
  - recessive = 25 percent,
  - dominant = 25 percent,
  - additive = 25 percent, and
  - multiplicative = 25 percent.

We acknowledge that this distribution of MOI traits does not represent how inheritance traits are distributed in humans. The Online Mendelian Inheritance in Man (OMIM)[7] provides the best source of information on the MOI distribution (Table 4.1). However, OMIM is disproportionally populated by genes linked to single Mendelian disorders. Therefore, genes associated with multifactorial disorders are under-represented in OMIM. Because polygene influences are assumed to be a major source of additive and multiplicative SNP behavior, the distribution in Table 4.1 is likely biased. Accordingly, we populated SNPs in our data with equal MOI representation and acknowledge that it does not represent the true distribution.

Table 4.1. Distribution of genes in Online Mendelian Inheritance in Man, by mode of inheritance (MOI)

MOI	Frequency
Autosomal Dominant	3,805
Autosomal Additive	12
Autosomal Multiplicative	21
Autosomal Recessive	3,775

The three optimal MOI specific methods are the three variations of the CA trend test described in Zheng and Gastwirth.[8] We also included a fourth, commonly used individual method, the 2-degrees-of-freedom (2*df*) genotype association test. Using the notation in Table 4.2 to define the 2 × 3 table of case-control counts stratified by genotype, a one-tailed test statistic ($T^2(x)$) for the three variations of the CA trend methods is defined as:

$$T^2(x) = n \frac{[\Sigma_{0,1,2} \ \{ x_i \, (s \, r_i - \ r \, s_i)\}]^2}{[r \, s \, (\Sigma_{0,1,2} \ \ n \, \{x_i \, x_i \, n_i \} - \{\Sigma_{0,1,2} \ \ (x_i \, n_i)^2 \})]} . \qquad (4.1)$$

The values represented in equations (4.1) and (4.2) are shown in Table 4.2, and the value of x_i defines the specific test $x_0 = 0$, $x_2 = 1$ and $x_1 = \{0$ – recessive, .5 – additive, 1 – dominant).

Table 4.2. Terms defined in equations (4.1) and (4.2)

	AA	Aa	aa	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

AA = major genotype; **Aa** = heterozygote genotype; **aa** = minor genotype.

Under the null hypothesis of no association, $T^2(x)$ has an asymptomatic χ^2 distribution with 1 degree of freedom. Please note that the power of the trend tests may be affected by the variance estimator used. In particular, the usual method of combining both cases and controls is not an asymptotically unbiased estimator of the null variance when the alternative is true. The authors note that at least two different estimates of the null variance are available, which are consistent under both the null and alternative hypotheses. In our calculations, we use a one-sided test.

As an alternative to equation (4.1), it is also possible to use a normally distributed test statistic, per Li et al.[2]:

$$N(x) = n^{1/2} \frac{[\Sigma_{0,1,2} \{ x_i (s r_i - r s_i)\}]}{[r s (n \Sigma_{0,1,2} \{x_i x_i n_i\} - [\Sigma_{0,1,2} \{x_i n_i\}^2]^{1/2})]} . \quad (4.2)$$

Under the null hypothesis of no association, N(x) has an asymptotic normal distribution N(0,1), which suggests a one-tailed test because the synthetic data assumes that the minor allele conveys the risk of phenotype.

We used eight sample size assumptions with equal numbers of cases and controls to perform our analysis, with *N* defined as the number of cases: = 100, 250, 500, 1,000, 2,000, 4,000, 8,000, or 9,500. We estimate statistical power by statistical method and *N* using a significance threshold of $\alpha = 10^{-7}$. In a GWAS, researchers usually perform a single marker analysis as a starting point to identify SNPs for additional and more comprehensive analysis. This initial pass creates a large number of statistical tests as well as a high potential for false-positive predictions, which has caused researchers to perceive the need for a very low threshold. Accordingly, some studies have used type I threshold levels on the order of 10^{-7}.[9-11]

The multitest statistical methods we used in our comparisons are:

1. The Bonferroni (BON) method, shown in Holm,[12] which is a simple form of the Bonferroni correction results which uses *n* methods to test for an association outcome. The correction involves dividing the alpha level by *n*. For example, if the association of a given SNP involves using three different statistical methods, the corrected alpha level (α) would be α/3. This would ensure that the overall chance of making a Type I error is still less than α.
2. The MAX method from Li et al.[2] that departs from the Bonferroni method. Bonferroni assumes that the individual tests are mutually independent; while Li et al. assume that the individual tests are correlated and incorporate an approximation to the joint distributions.

Results: Statistical Method Assessment

Table 4.3 presents power estimates by statistical methods and sample size and is based on a fixed alpha threshold ($\alpha = 10^{-7}$). All tests are one-sided, and the tests included in this table are:

- The additive X2 version of the CA (CA-A) test, which was the best method for both additive and multiplicative gene models but was not particularly effective when applied to recessive MOI data
- The Bonferroni test (BON), which combines the X2 version of the recessive, additive, and dominant MOI specific tests (CA-R, CA-A, CA-D) and improves on the test performance of the three individual tests when the MOI gene model is not known
- The MAX test due to Li et al.,[2] which combines the normal version of the CA-R, CA-A, and CA-D tests to improve on the test performance of the three individual tests if the MOI gene model is not known
- The dominant X2 version of the CA test (CA-D), which was the best method for the dominant gene models
- The recessive X2 version of the CA test (CA-R), which was the best method for recessive gene models but the least effective when applied to nonrecessive MOI data
- The 2*df* genotype test (2*df*-g), which is never an optimal method for any of the scenarios we examined; in every scenario, a more powerful alternative can be identified (see Table 4.3)

Table 4.3. Power results, by statistical method and number of cases: additive mode of inheritance data

N	CA-A	BON	MAX	CA-D	CA-R	2*df*-G
100	2.26*	1.73	1.91**	0.62	0.46	0.00
250	13.47*	12.22	12.25**	9.46	5.56	2.41
500	30.00*	28.12**	28.06	24.19	12.63	10.48
1,000	49.91*	48.31**	48.26	45.72	25.11	25.21
2,000	68.80*	67.40**	67.31	64.72	41.08	46.35
4,000	83.00*	81.86	81.94**	80.25	58.24	65.96
8,000	94.30*	93.69	93.72**	92.58	72.36	80.52
9,500	96.07*	95.63**	95.46	95.01	75.66	83.90

2*df*-G = 2-degrees-of-freedom genotype association test; BON = Bonferroni test; CA-A = autosomal additive Cochran-Armitage test; CA-D = autosomal dominant Cochran-Armitage test; CA-R = autosomal recessive Cochran-Armitage test; MAX = MAX combined test.

* Best power score ** Second best power score

Our results indicate that the best method in terms of statistical power is CA-A, but that little is lost if the BON or MAX method is used instead. Also there is little difference between the BON and MAX methods. Similarly, the results in Tables 4.4, 4.5, and 4.6 indicate that the best method in terms of statistical power for identifying dominant MOI loci is CA-D, and CA-R for recessive MOI loci. For multiplicative MOI loci, the best method is CA-A. In all four scenarios, little is lost if the MAX or BON method is used as a replacement.

Table 4.4. Power results, by statistical method: dominant mode of inheritance gene data

N	CA-A	BON	MAX	CA-D	CA-R	2*df*-G
100	0.05	0.07	0.40**	0.13*	0.00	0.00
250	2.09	2.94**	2.88	3.79*	0.00	0.00
500	12.89	15.01	15.02**	16.52*	0.01	1.21
1,000	32.75	34.75	34.84**	36.94*	0.19	12.79
2,000	54.80	56.33	56.55**	57.97*	2.42	32.75
4,000	71.01	73.48**	73.46	74.90*	11.24	54.50
8,000	85.15	86.95	87.18**	88.09*	26.71	71.98
9,500	88.27	90.15**	90.10	91.23*	31.55	76.39

2*df*-G = 2-degrees-of-freedom genotype association test; BON = Bonferroni test; CA-A = autosomal additive Cochran-Armitage test; CA-D = autosomal dominant Cochran-Armitage test; CA-R = autosomal recessive Cochran-Armitage test; MAX = MAX combined test.

* Best power score ** Second best power score

Table 4.5. Power results, by statistical method: recessive mode of inheritance gene data

N	CA-A	BON	MAX	CA-D	CA-R	2*df*-G
100	0.00	0.00	0.38**	0.00	0.00*	0.00
250	0.02	0.11	0.68**	0.00	0.19*	0.00
500	0.42	1.70	1.77**	0.00	2.11*	0.01
1,000	2.71	7.88	7.89**	0.01	8.95*	1.04
2,000	9.97	19.91	19.99**	0.13	21.40*	6.50
4,000	21.84	34.29	34.39**	1.50	35.95*	18.45
8,000	34.94	49.35	49.67**	7.17	51.01*	33.14
9,500	38.04	53.07	53.11**	9.32	54.62*	36.73

2*df*-G = 2-degrees-of-freedom genotype association test; BON = Bonferroni test; CA-A = autosomal additive Cochran-Armitage test; CA-D = autosomal dominant Cochran-Armitage test; CA-R = autosomal recessive Cochran-Armitage test; MAX = MAX combined test.

* Best power score ** Second best power score

Table 4.6. Power results, by statistical method: multiplicative mode of inheritance gene data

N	CA-A	BON	MAX	CA-D	CA-R	2*df*-G
100	4.00*	3.35	3.67**	0.98	1.65	0.01
250	17.16*	15.60	15.71**	11.07	8.82	4.74
500	35.65*	33.81**	33.77	27.25	19.4	12.50
1,000	53.84*	52.08**	51.78	48.11	33.65	30.72
2,000	71.59*	70.31**	70.27	66.44	49.68	50.69
4,000	85.12*	84.03**	83.96	81.40	64.32	67.66
8,000	95.10*	94.58**	94.58**	93.36	77.30	82.84
9,500	96.36*	95.99	96.08**	95.30	80.27	86.42

2*df*-G = 2-degrees-of-freedom genotype association test; BON = Bonferroni test; CA-A = autosomal additive Cochran-Armitage test; CA-D = autosomal dominant Cochran-Armitage test; CA-R = autosomal recessive Cochran-Armitage test; MAX = MAX combined test.

* Best power score ** Second best power score

However, regardless of MOI, power is lost if we use the CA-A method. Many researchers use an additive model as the initial GWAS pass. What if the locus in question is recessive or dominant? Table 4.7 indicates that although the CA-A method is the optimal choice (by as much as 2 percent), if the MOI of the locus is additive or multiplicative, there is a risk of more than 2 percent power loss if the locus MOI is dominant and as much as 15 percent loss if the MOI of the locus is recessive.

Table 4.7. Power results, by CA-A and MAX methods, for different MOI gene models

N	CA-A (D)	MAX	Loss using CA-A	CA-A (R)	MAX	Loss using CA-A
100	0.05	0.40	.35	0.00	0.38	.38
250	1.45	2.88	1.43	0.02	0.68	.66
500	12.89	15.02	2.13	0.42	1.77	1.35
1,000	32.75	34.84	2.09	2.71	7.89	5.18
2,000	54.80	56.55	1.75	9.97	19.99	10.02
4,000	71.01	73.46	2.45	21.84	34.39	13.55
8,000	85.15	87.18	2.03	34.94	49.67	14.73
9,500	88.27	90.10	1.83	38.04	53.11	15.07

CA-A (D) = Cochran-Armitage method, additive model version applied to dominant mode of inheritance (MOI) single nucleotide polymorphism (SNP) data; CA-A (R) = Cochran-Armitage method, additive model version applied to recessive MOI SNP data; MAX = MAX combined test.

If we knew the distribution of the MOI property, we could assess the overall risk of using an additive method such as CA-A for GWAS. However, without a reliable estimate, researchers should exercise caution and apply a procedure that limits the risk of incorrectly assessing the MOI inherent to the locus-inducing diagnosis.

Discussion

In the literature, many statistical methods that have been used to perform GWAS assume a MOI specific hypothesis. Our results confirm the work of many others.[9] In the context of a single-marker scenario, the best method for predicting associations in recessive SNPs is the CA-R method; the best method for dominant MOI SNPs is the CA-D method; and the best method for additive and multiplicative SNPs is the CA-A method.

We also show that the 2*df* genotype method used in many studies—for example, the method used by Schymick et al.[6]—is never optimal because there are always other methods that provide greater statistical power. This statement holds regardless of whether the MOI is known *a priori* or not. We also show that in the context of a general method to use in the initial GWAS pass, researchers may encounter adverse consequences if, for example, the MOI of the operating locus is not consistent with the assumption employed by the statistical method used. Therefore, 2*df* appears to be inappropriate to use for GWAS under any circumstances.

Consequently, we examined the possibility of employing an alternative procedure that incorporates the three core tests defined above into two multitest procedures: BON, a Bonferroni corrected procedure, and the MAX test procedure developed by Li et al.[2] Both methods are a composite of three separate tests (additive, dominant, and recessive models). These procedures are opposites in that they assume different underlying distributions of the test statistics. The MAX method assumes that the three tests have dependencies that can be accounted for, whereas the Bonferroni method assumes that the three tests are mutually independent. We note that despite these differences, the two methods produce very similar power profiles.

We generated our results using 1,000 replicates per parameter combination. Our standard error estimate of power varies from .262 to .315. Consequently, our 95 percent confidence interval around the mean will be approximately plus or minus .019. Although we recognize that a larger number of replicates will improve power precision, we believe that our conclusion would remain as stated.

In summary, our results lead us to recommend that researchers use a statistical methodology that combines the results of individual MOI-based core tests as a statistical method for conducting GWAS rather than a 2*df* test. Combining individual methods and comparing the individual and combined results may help identify the MOI character of the gene. The actual process of combining the core tests into a single operational test can be done in a number of ways, all of which produce very similar statistical power profiles.

Chapter References

1. Cooley P, Clark R, Folsom R, et al. Genetic inheritance and genome wide association statistical test performance. J Proteomics Bioinform. 2010;3(12):321-325.
2. Li Q, Zheng G, Li Z, et al. Efficient approximation of P-value of the maximum of correlated tests, with applications to genome-wide association studies. Ann Hum Genet. 2008;72(Pt 3):397-406.
3. Kuo CL, Feingold E. What's the best statistic for a simple test of genetic association in a case-control study? Genet Epidemiol. 2010;34(3):246-253.
4. Sasieni PD. From genotypes to genes: doubling the sample size. Biometrics. 1997;53(4):1253-61.

5. Freidlin B, Zheng G, Li Z, et al. Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered. 2002;53(3):146-52.

6. Schymick JC, Scholz SW, Fung HC, et al. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol. 2007;6(4):322-8.

7. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine. Online Medelian Inheritance in Man (OMIM). 2016 [cited 2016 Feb 11]; Available from: http://www.ncbi.nlm.nih.gov/omim

8. Zheng G, Gastwirth JL. On estimation of the variance in Cochran-Armitage trend tests for genetic association using case-control studies. Stat Med. 2006;25(18):3150-9.

9. Iles MM. Effect of mode of inheritance when calculating the power of a transmission/disequilibrium test study. Hum Hered. 2002;53(3):153-7.

10. Ziegler A, Konig IR, Thompson JR. Biostatistical aspects of genome-wide association studies. Biom J. 2008;50(1):8-28.

11. van Es MA, van Vught PW, Blauw HM, et al. Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. Nat Genet. 2008;40(1):29-31.

12. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6:65-70.

CHAPTER 5

The Influence of Errors Inherent in Genome-Wide Association Studies (GWAS) in Relation to Single-Gene Models

Philip Chester Cooley, Robert F. Clark, and Grier Page

Overview

The influence of genotype and diagnosis errors present in genome-wide association studies (GWAS) was assessed by analyzing a synthetic gene data set incorporating factors known to influence association measurement. Monte Carlo methods were used to generate the synthetic gene data that incorporated factors that influenced including gene inheritance, relative risk levels, disease penetrance, genotype distribution, sample size, as well as the two error factors that are the focus of this study. The resulting data set provides a truth set for assessing statistical method performance and association sensitivity.

Our results quantify the relationship between genotype and diagnosis error measures and statistical power loss. The connection between these relationships are understood, but we document their extent. Our results also demonstrate that for low-risk nonrecessive loci, sample sizes in the range of 1,000 to 2,000 cases will achieve 80 percent power thresholds for type-I error levels of 10^{-8}, even with realistic genotype and phenotype error assumptions. Nevertheless, increasing sample size is a viable method of compensating for power loss caused by genotype and diagnosis errors. Our estimates indicate that sample sizes should be increased by 20 to 40 percent, depending on the gene inheritance model assumed.

Chapter 5 is based on a study that was published in the *Journal of Proteomics & Bioinformatics*.[1]
 Chapter 3 of the current publication describes the generation of the synthetic gene database. This section was removed from Chapter 5. The analysis of the gene data has not changed.

Introduction

More than 2,306 human GWAS have examined more than 1,000 diseases and traits and found more than 1,200 single nucleotide polymorphism (SNP) associations.[2] With improved genotyping technologies and the growing number of available markers, case-control GWAS have become a key tool for investigating complex diseases. Because GWAS have become a standard primary investigative tool, researchers need to be aware of how errors influence their studies and how to overcome or compensate for them. The initial step in a GWAS is to apply univariate statistical tests for each SNP in the data set. Applying the tests is statistically straightforward and uses several standard approaches (e.g., χ^2 tests, regression methods).

Studies on the consequences of genotype error have led to a modest number of investigations in the statistical genetics literature. Gordon and colleagues investigated the effects of three published models of genotyping errors on the 2*df* genotype χ^2 test.[3] In another study, Gordon and colleagues described a statistical power calculator (PAWE-3D) that produces power and sample size calculations that can support study designs for GWAS and compute power and/or sample size requirements for a specified significance level.[4] Zheng and Tian, as well as Edwards and colleagues contributed to the development of PAWE.[5,6] Gordon and colleagues further analyzed the influence of both random phenotype and genotype misclassification errors on statistical power contrasting the Cochran-Armitage additive test (CA-A) with the 2-degrees-of-freedom (2*df*) genotype test and concluded that the CA-A is more powerful.[7]

Ahn and colleagues addressed the effect of different types of genotyping errors on statistical power in GWAS.[8] Although their prior work focused on non-differential genotype error rates, this study considered errors in each of the three bi-allelic genotypes differentially. The methods were based on a Taylor-series expansion of a noncentrality parameter of the asymptotic distribution of the trend test. In a follow-up study, Ahn and colleagues extended their work by developing a closed form analytic procedure for both the 2*df* genotype and the CA-A tests.[9] They reported that misclassifying the heterozygote genotype is particularly detrimental when using the Cochran-Armitage recessive trend test (CA-R) on data from a recessive mode of inheritance (MOI) model.

Although the accuracy of the genotyping process has improved, data errors still occur. Hao and colleagues reported an overall 0.5 percent error rate imputation process, but they also reported a 2 percent error rate in

underrepresented subpopulations.[10] Miclaus and colleagues examined genotype calling algorithms on HapMap samples and found that different algorithms can produce genotyping errors that influence downstream genotype calls.[11] They reported a 2 to 3 percent error estimate attributable to the genotype-calling algorithm. Laurie and colleagues estimated genotyping error rates from duplicate sample discordance rates from addiction and lung cancer projects genotyped on Illumina Human1Mv1_c and HumanHap550-2v3_b arrays by the Center for Inherited Disease Research.[12] The investigators calculated genotyping error rates on the order of 10^{-4}, which corresponds to mean completion call rates of 99.7 and 99.8 percent, respectively, for the two projects. If study samples are not duplicated, as in the type-2 diabetes project, but with multiple replicates of the HapMap control sample, discordant rates of $1\text{-}4 \times 10^{-3}$ lead to completion rates of 99.6 to 99.7 percent.

Phenotypic misclassification errors are also a source of bias and can reduce the power to detect a statistical association between a phenotype and a specific allele.[13,14] Edwards and colleagues presented a quantification of the effect of phenotypic error on power and sample size calculations for case-control genetic association studies between a marker locus and a disease phenotype.[6] Their process is specific to the standard case-control method used commonly in epidemiology, from which they develop a process that quantifies power loss and minimum sample size requirements in the presence of phenotypic errors. Barendse described the effect of investigator measurement-error on the phenotypes—the error was significant when looking at quantitative traits.[14] When the traits were coded as affected or unaffected, the error effect sizably decreased (from 14.5 to 5.3 percent). For many diseases, the interrater agreement for disease diagnosis can be quite low. For example, the sensitivity and specificity of predeath Alzheimer's diagnosis with most post mortem autopsies can be as low as 0.83 and 0.84 respectively.[15]

Gene traits that influence prediction accuracy have been also reported in other studies. For example, Sasieni's work as well as that of Freidlin and colleagues has demonstrated that the phenotype MOI model was a major influence on association prediction accuracy.[16,17]

To provide additional insight into the influence of genotype and diagnosis errors affecting the accuracy of the phenotype measure in a GWAS, we ran simulations with synthetically generated data. We focused on assessing the impact on statistical power caused by the influence of these two often-overlooked errors. Our simulations demonstrated that genotype (even at low

error rates) and phenotype (diagnosis) errors produce substantial power losses for all MOIs, with significant power losses for recessive MOIs. Because GWAS involving recessive loci have additional power requirements relative to other MOI types, researchers need to address these requirements in developing appropriate sample sizes for their studies.

Methods

Our approach is based entirely on a simulation framework. Chapter 3 describes in detail the data generation method that was used to produce our synthetic data and how we used these data to assess the influence of errors on statistical power loss.

We developed our assessments by analyzing a data set of synthetic gene data that incorporates factors that we know influence association measurements in GWAS. These include phenotypic errors (i.e., those caused by improper disease diagnosis) and genotype errors (e.g., those caused by incorrect genotype calls). We employed Monte Carlo methods to generate simulated gene data that we analyzed to assess the influence of the individual factors on statistical power in the context of GWAS. There are two advantages to using simulated data. First, the association-affecting factors are isolated and can be linked to the affecting locus. Second, we can choose any specific statistical method to perform the association assessment. The simulated data set provides a truth set for assessing the role of statistical methods on association sensitivity and highlights the particular role of errors in disease diagnosis and incorrect genotype assignments.

Results

Data Set Summary

Using the methods described in Chapter 3, we generated a synthetic gene simulated data set with the following characteristics:

- The proportion of cases (controls) that are major homozygotes = 50.3 (63.0) percent.
- The proportion of cases (controls) that are heterozygotes = 39.2 (31.3) percent.
- The proportion of cases (controls) that are minor homozygotes = 10.5 (5.7) percent.
- With MOI distribution:

- recessive = 25 percent,
- dominant = 25 percent,
- additive = 25 percent, and
- multiplicative = 25 percent.

Although this distribution of MOI traits does not represent a "true" distribution, we currently know of no accurate way to obtain such a distribution. Consequently, although we gave each of the four MOI traits equal representation in the simulated data, we confined our examinations to within-MOI assessments.

Factor Assessment

To check the influence of the factors (relative risk, penetrance, MOI, genotype error rates, phenotype error rates) built into the data for each experiment, we fitted the three distinct models to the four MOI subsets of the data: that is, if the index *i* represents the *i*th (of 12) experiments, where the model can be described as:

$$-\log_{10}(p_i) = \beta 0 + \beta 1 \times n_i + \beta 2 \times dp_i + \beta 3 \times ErrP_i + \beta 4 \times ErrG_i + \beta 5 \times \Phi_i + \beta 6 \times G_i \quad (5.1)$$

With a single exception, the signs of the estimated phenotype error term (β3) and genotype error term (β4) coefficients were consistent and the magnitudes of the estimated coefficients similar. All estimates were significant at the 10^{-7} level, suggesting that the factors incorporated in the simulated data are all significant.

Error Analyses

To simulate the association estimation process in a GWAS experiment, we applied three variations of the Cochran-Armitage (CA) trend test to each of the 1,000 replicates of the 3,456 possible data subsets. Each of the variations of the CA tests used a distinct genotype score vector: [0,0,1] for recessive, [0,0.5,1] for additive and [0,1,1] for dominant. We applied each of the tests to all of the replicates in each of the data subsets. This process allowed us to show that the optimal strategy for maximizing statistical power is MOI specific. This strategy posits that the recessive version (CA-R) be used to estimate associations involving recessive loci data, the dominant version (CA-D) for the dominant loci data, and the additive version (CA-A) for both additive and multiplicative

loci data. This strategy is cited by others for single-gene models.[18] Cooley and colleagues provided a similar assessment and identified a multiple test strategy that combines the three tests into an overall score that has merit if the MOI of the causative loci is not known.[19] For the assessment in this study, we assumed that the MOI is known and selected the best statistical method to measure the association. Consequently, our results tended to be optimistic.

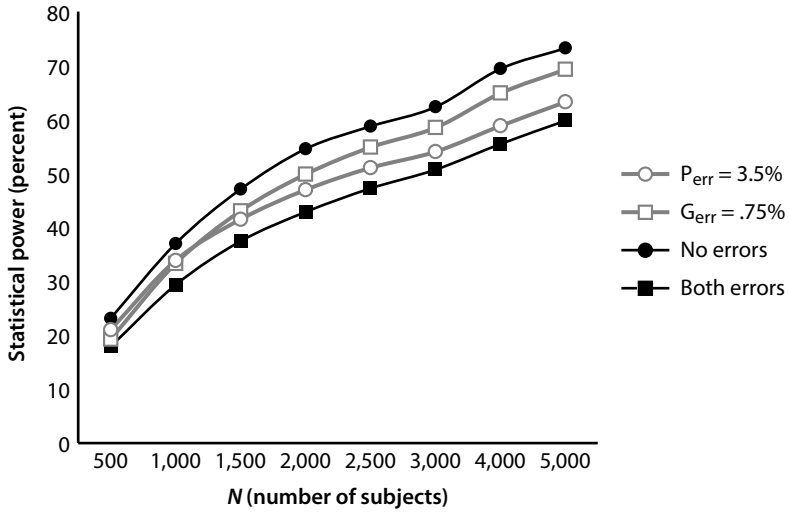
Error rates of 0, 2, and 5 percent are incorporated into the simulated data for the phenotype and 0, .5, and 1 percent for the genotype. Our approach combines the three risk levels (mean risk = 1.3), three penetrance levels (mean penetrance = 0.4), and groups the data into a "with error" (mean phenotype error = 3.5 percent, and mean genotype error 0.75 percent) and "without error" strata. Also, we also stratified the analysis by MOI. Figures 5.1a through 5.4b identify the four MOI-specific results. Each figure includes a 0.75 percent genotype error curve, a 3.5 percent phenotype error curve, a curve that includes both error sources, and a curve generated without either source of error. Figure 5.1 presents the recessive loci analysis. The impact of a 0.75 percent average genotype error rate and a 3.5 percent average diagnosis error rate with respect to power loss for recessive loci is nontrivial. However each profile exhibits distinct behavior. The effect of the phenotype error increases with N and peaks at $N = 4{,}000$ cases, whereas the genotype error effect is constant across all N. Also observed at the peak is a genotype impact of 6.04 percent power loss per 1.0 percent genotype error and a power loss of 3.03 percent power loss per 1.0 percent phenotype error. Note that with $\alpha < 10^{-8}$ as the significance threshold, an 80 percent power target is far from being realized even with $N = 5{,}000$ cases and controls.

Figures 5.2, 5.3, and 5.4 present the error effects of the dominant, additive, and multiplicative loci respectively. All three figures indicate that power loss is nontrivial for the MOI categories they represent but that the effect is substantially less than recessive modes.

The pattern of the power profile for dominant loci is in sharp contrast to the recessive loci profile. The diagnosis error pattern is constant across N for dominant loci—the recessive loci show an increasing pattern. The genotype patterns are also different. As N increases, the power differences decline for the dominant loci, whereas the patterns are constant for recessive loci.

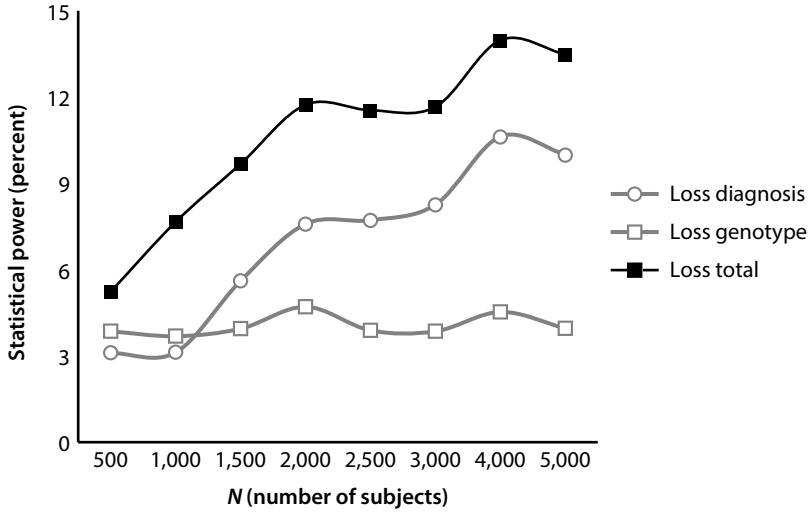
The additive and the multiplicative loci show similar error profiles. The impact of a 3.5 percent diagnosis error and the impact of a 0.75 percent genotype error have a similar quantitative impact. Both have a declining power loss as N increases.

Figure 5.1a. The impact of genotype and diagnosis errors on power: recessive loci



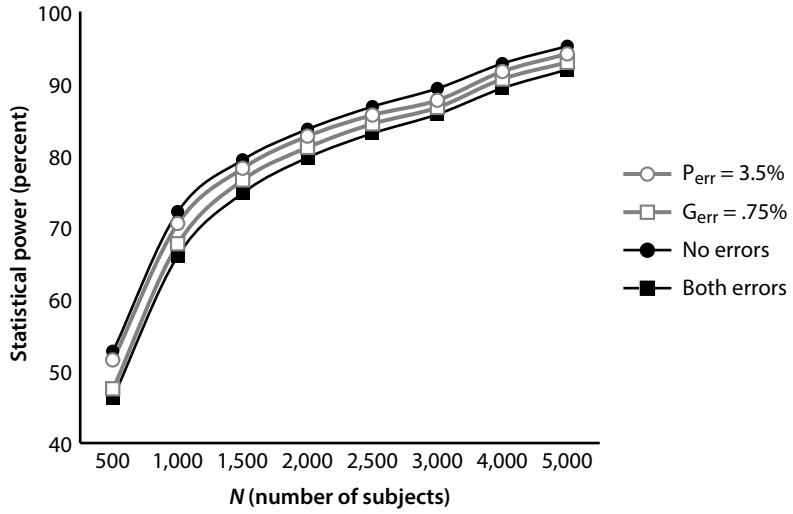
G_{err} = genotype error level
P_{err} = phenotype (diagnosis) error level
Both errors = both genotype and phenotype errors
No errors = neither error

Figure 5.1b. Power loss: total, genotype, and diagnosis: recessive loci



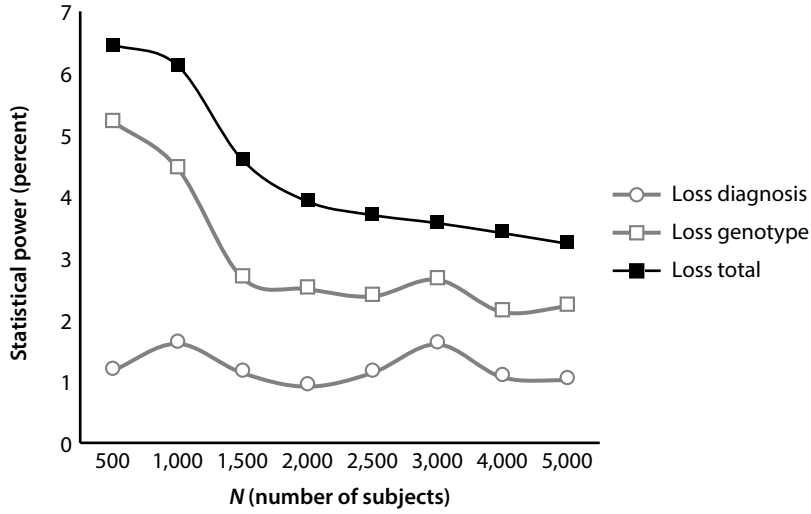
Loss genotype = power difference between G_{err} = .75 and no errors assumption.
Loss phenotype (diagnosis) = power difference between P_{err} = 3.5 and no errors assumption.
Loss total = power difference between both P_{err} = 3.5 & G_{err} = .75 and no errors assumption.

Figure 5.2a. The impact of genotype and diagnosis errors on power: dominant loci



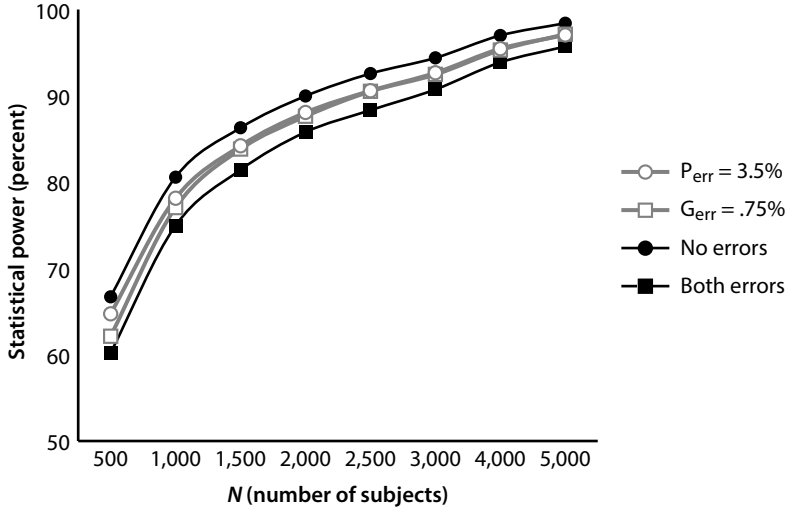
G_{err} = genotype error level
P_{err} = phenotype (diagnosis) error level
Both errors = both genotype and phenotype errors
No errors = neither error

Figure 5.2b. Power loss: total, genotype, and diagnosis: dominant loci



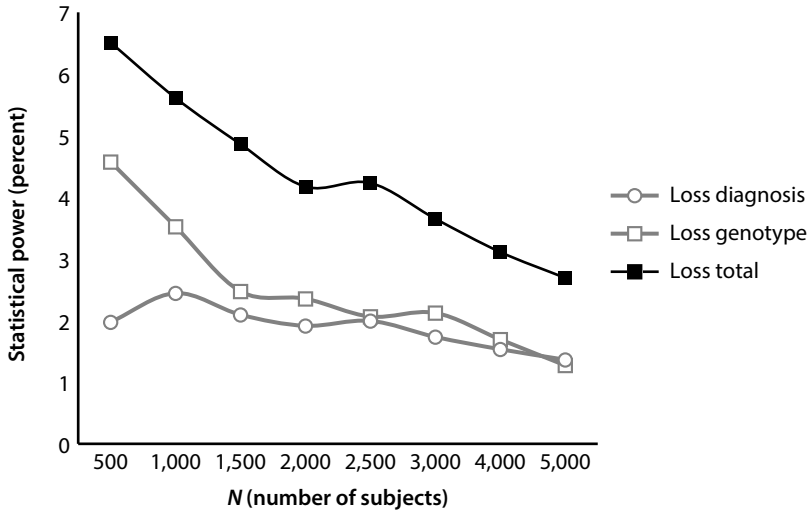
Loss genotype = power difference between G_{err} = .75 and no errors assumption.
Loss phenotype (diagnosis) = power difference between P_{err} = 3.5 and no errors assumption.
Loss total = power difference between both P_{err} = 3.5 & G_{err} = .75 and no errors assumption.

Figure 5.3a. The impact of genotype and diagnosis errors on power: additive loci



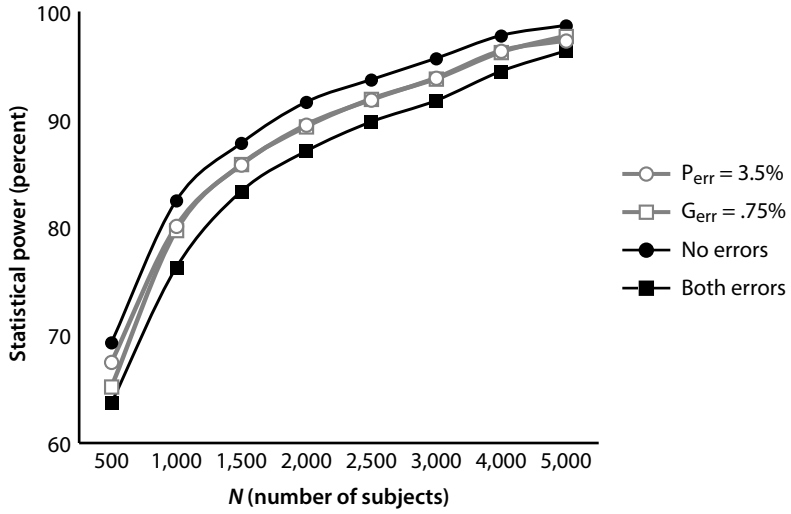
G_{err} = genotype error level
P_{err} = phenotype (diagnosis) error level
Both errors = both genotype and phenotype errors
No errors = neither error

Figure 5.3b. Power loss: total, genotype, and diagnosis: additive loci



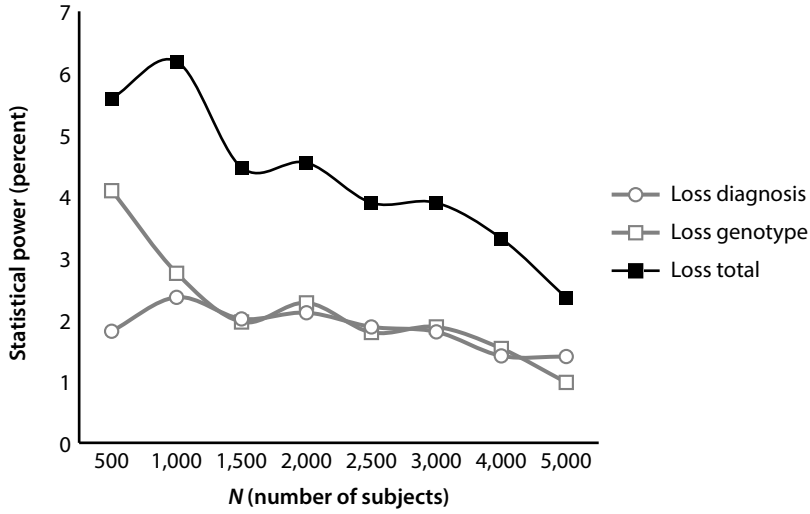
Loss genotype = power difference between G_{err} = .75 and no errors assumption.
Loss phenotype (diagnosis) = power difference between P_{err} = 3.5 and no errors assumption.
Loss total = power difference between both P_{err} = 3.5 & G_{err} = .75 and no errors assumption.

Figure 5.4a. The impact of genotype and diagnosis errors on power: multiplicative loci



G_{err} = genotype error level
P_{err} = phenotype (diagnosis) error level
Both errors = both genotype and phenotype errors
No errors = neither error

Figure 5.4b. Power loss: total, genotype, and diagnosis: multiplicative loci



Loss genotype = power difference between G_{err} = .75 and no errors assumption.
Loss phenotype (diagnosis) = power difference between P_{err} = 3.5 and no errors assumption.
Loss total = power difference between both P_{err} = 3.5 & G_{err} = .75 and no errors assumption.

In summary, the four figures indicate that the genotype error versus the diagnosis errors effects vary by MOI. For the recessive MOI, a 3.5 percent diagnosis error has a larger impact than a 0.75 percent genotype error. This result is reversed in the dominant MOI scenarios (Figure 2). The additive and the multiplicative MOI scenarios represented in Figures 3 and 4 indicate that a 0.75 percent genotype error is comparable in effect to a 3.5 percent diagnosis error with respect to power loss.

These results are summarized in Table 5.1, which displays the power loss for the smallest sample size ($N = 500$ cases) and the largest sample size ($N = 5{,}000$ cases). For example, row R (recessive) of Table 5.1 illustrates that error loss due to genotype errors at $N = 500$ and $N = 5{,}000$ is flat, but that error loss due to diagnosis error increases dramatically from $N = 500$ to $N = 5{,}000$ and dominates the total error profile. The power loss pattern changes for the other three MOIs where error loss patterns for both genotype and diagnosis sources decline from $N = 500$ to $N = 5{,}000$.

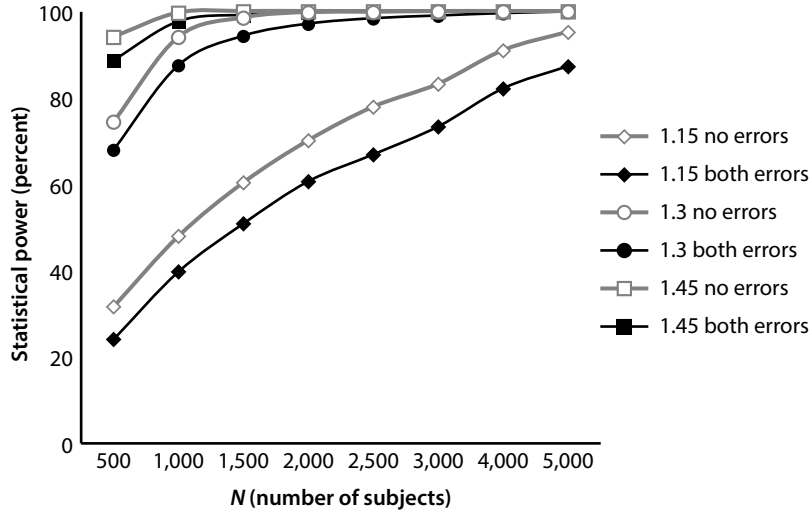
Table 5.1. Maximum power loss due to genotype or diagnosis error

	Genotype Errors		Diagnosis Errors		Both Errors	
MOI	***N* = 500**	***N* = 5,000**	***N* = 500**	***N* = 5,000**	***N* = 500**	***N* = 5,000**
R	3.86	3.96	3.11	9.98	5.23	13.45
D	5.23	3.43	1.21	1.06	6.46	3.26
A	4.57	1.27	1.97	1.36	6.51	2.69
M	4.09	0.98	1.81	1.40	5.58	2.35

A = additive; D = dominant; MOI = mode of inheritance; M = multiplicative; R = recessive.

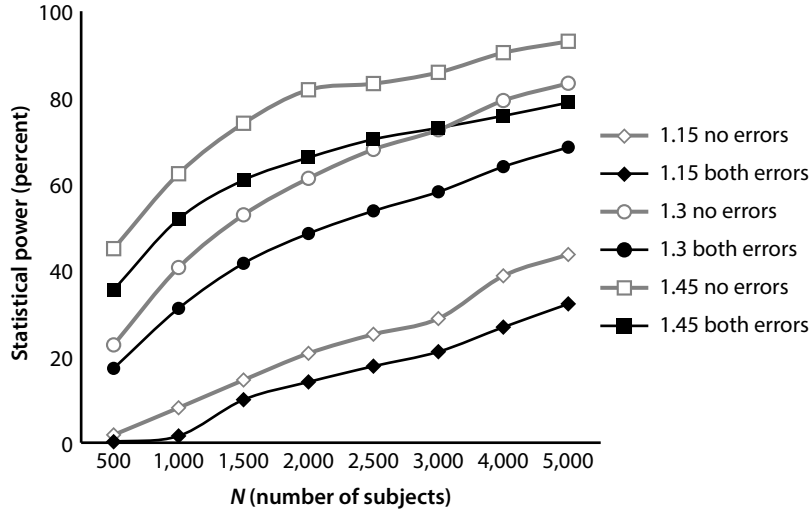
We also examined the simultaneous influence of relative risk and error effects on statistical power. As above, we analyzed our data set using all three penetrance levels (mean penetrance = 0.4), but we also stratified the curves by the low- (1.15), medium- (1.3) and high-risk (1.45) categories. Figure 5.5 displays the combined (genotype plus diagnosis) error effects for the three risk categories using the CA-A method applied to the additive MOI data. Similar curves can be generated for the dominant and multiplicative scenarios. This figure suggests that, for additive inheritance scenarios, researchers can predict associations in the context of GWAS with a type-I error threshold of $\alpha < 10^{-8}$ and still achieve a power level greater than 80 percent. This statement applies to low-risk loci even when diagnosis errors are 3.5 percent and genotype errors are 0.75 percent.

Figure 5.5. Genotype error: recessive mode of inheritance, by risk level and *N*



Loss genotype = power difference between G_{err} = .75 and no errors assumption.
Loss phenotype (diagnosis) = power difference between P_{err} = 3.5 and no errors assumption.
Loss total = power difference between both P_{err} = 3.5 & G_{err} = .75 and no errors assumption.

Figure 5.6. Phenotype error: recessive mode of inheritance, by risk level and *N*



Loss genotype = power difference between G_{err} = .75 and no errors assumption.
Loss phenotype (diagnosis) = power difference between P_{err} = 3.5 and no errors assumption.
Loss total = power difference between both P_{err} = 3.5 & G_{err} = .75 and no errors assumption.

Figure 5.6 shows the same results for the recessive scenario. In this recessive scenario, the likelihood of achieving an 80 percent power level is low and is only possible for high-risk loci in the absence of genotype and diagnosis error with a sample size *N* larger than attempted by our simulation experiments.

Summary/Discussion

We examined the influence of genotype and diagnosis errors that affect the accuracy of association predictions in a GWAS and focused on assessing the effect on statistical power loss caused by the influence of these two sources of error. Our findings are MOI specific and indicate that both sources of error can adversely affect power levels. This outcome is more pronounced for recessive MOI and low-risk loci, which is common knowledge. What our study shows is that the error magnitude depends on a variety of factors in addition to MOI, especially relative risk and sample size; our study quantifies this magnitude and indicates the significance of this impact. This loss can be compensated for by increasing sample sizes. Gordon and colleagues reported that a 1 percent increase in genotype error rates requires an increase in sample size of 2 to 8 percent, which they also noted depends on the MOI scenario.[3] Our estimates are much higher than those reported by Gordon and colleagues and are based on achieving a power threshold of 80 percent. Using the additive model, results at $N = 1{,}000$ (assuming no genotype errors) exceeds the 80 percent threshold (80.6 percent). Introducing a 1 percent genotype error, power drops to 75 percent. An additional 405 cases are needed to compensate for this loss to restore an 80.6 percent power level, which is a 40.5 percent increase in sample size. Please note that we are not suggesting that 1 percent error is standard operating procedure. In fact, genotype errors are improving with the introduction of each new technology,and currently are likely less than 0.5 percent. Table 5.2 presents these results for all MOIs for both genotype and diagnosis errors.

Table 5.2. Percent sample size increase to restore power caused by a 1 percent genotype or diagnosis error

MOI	Genotype percentage	Diagnosis percentage
R	57.2	35.9
D	40.1	19.7
A	40.5	20.9
M	40.2	18.9

A = additive; D = dominant; MOI = mode of inheritance; M = multiplicative; R = recessive.

In summary, our results quantify the relationship between genotype and diagnosis error measures and statistical power loss. These relationships are understood, but we document their extent. Our results also assume that we know the MOI of the locus being analyzed; therefore, our results will understate the true power loss and the compensating sample size increases. Our results also demonstrate that for low-risk nonrecessive loci, sample sizes in the range of 1,000–2,000 cases will achieve 80 percent power thresholds for type–I error levels of 10^{-8} even with realistic genotype and phenotype error assumptions.

However, the recessive loci model remains problematic. Desirable power thresholds for moderate risk levels can only be realized with sample sizes in the tens of thousands, further complicated by accounting for power loss as a result of genotype and diagnosis errors.

Chapter References

1. Cooley P, Clark RF, Page G. The influence of errors inherent in genome wide association studies (GWAS) in relation to single gene models. J Proteomics Bioinform. 2011;4:138-144.
2. Burdett T, Hall P, Hasting E, et al. The NHGRI-EBI Catalog of published genome-wide association studies. 2015 [cited 2015 Nov 2]; Available from: www.ebi.ac.uk/gwas
3. Gordon D, Finch SJ, Nothnagel M, et al. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. Hum Hered. 2002;54(1):22-33.
4. Gordon D, Haynes C, Blumenfeld J, et al. PAWE-3D: visualizing power for association with error in case-control genetic studies of complex traits. Bioinformatics. 2005;21(20):3935-7.
5. Zheng G, Tian X. The impact of diagnostic error on testing genetic association in case-control studies. Stat Med. 2005;24(6):869-82.
6. Edwards BJ, Haynes C, Levenstien MA, et al. Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. BMC Genet. 2005;6:18.
7. Gordon D, Haynes C, Yang Y, et al. Linear trend tests for case-control genetic association that incorporate random phenotype and genotype misclassification error. Genet Epidemiol. 2007;31(8):853-70.

8. Ahn K, Haynes C, Kim W, et al. The effects of SNP genotyping errors on the power of the Cochran-Armitage linear trend test for case/control association studies. Ann Hum Genet. 2007;71(Pt 2):249-61.

9. Ahn K, Gordon D, Finch SJ. Increase of rejection rate in case-control studies with the differential genotyping error rates. Stat Appl Genet Mol Biol. 2009;8:Article25.

10. Hao K, Chudin E, McElwee J, et al. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. BMC Genet. 2009;10:27.

11. Miclaus K, Chierici M, Lambert C, et al. Variability in GWAS analysis: the impact of genotype calling algorithm inconsistencies. Pharmacogenomics J. 2010;10(4):324-35.

12. Laurie CC, Doheny KF, Mirel DB, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol. 2010;34(6):591-602.

13. Gordon D. Gene mapping: balance among quality, quantity and cost of data in the era of whole-genome mapping for complex disease. Euro J Hum Genet. 2006;14:1147-1148.

14. Barendse W. The effect of measurement error of phenotypes on genome wide association studies. BMC Genomics. 2011;12:232.

15. Blacker D, Albert MS, Bassett SS, et al. Reliability and validity of NINCDS-ADRDA criteria for Alzheimer's disease. The National Institute of Mental Health Genetics Initiative. Arch Neurol. 1994;51(12):1198-204.

16. Sasieni PD. From genotypes to genes: doubling the sample size. Biometrics. 1997;53(4):1253-61.

17. Freidlin B, Zheng G, Li Z, et al. Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered. 2002;53(3):146-52.

18. Kuo CL, Feingold E. What's the best statistic for a simple test of genetic association in a case-control study? Genet Epidemiol. 2010;34(3):246-53.

19. Cooley P, Clark R, Folsom R, et al. Genetic inheritance and genome wide association statistical test performance. J Proteomics Bioinform. 2010;3(12):321-325.

CHAPTER 6

Conducting Genome-Wide Association Studies (GWAS): Epistasis Scenarios

Philip Chester Cooley, Nathan Gaddis, Ralph E. Folsom, and Diane Wagener

Overview

In general, genome-wide association studies (GWAS) apply univariate statistical tests to each gene marker or single nucleotide polymorphism (SNP) as an initial step. This SNP-based test is statistically straightforward, and the core tests for assessing the associations are standard methods (e.g., χ^2 tests, regression) that have been studied outside of the GWAS context. Kuo and Feingold describe the most commonly used statistical methods that are applied to GWAS.[2] All tests cited in the chapter are single-locus tests. If the genetic inheritance properties are not known, we recommend combining two or more statistical tests.[3] In many cases, the SNPs associated with a disease are not located in a region of DNA that codes for a protein. Instead, they are located in the large noncoding regions between genes or in intron sequences, which are edited out of mRNAs prior to translation to proteins. These regions are presumably sequences of DNA that modify gene expression, but usually their functions are unknown.[4]

The popularity of the GWAS approach belies its simplicity and obscures the important issue of whether a single-gene model can illuminate the biosynthetic pathways of a phenotype. In the path leading from gene to trait, factors such as epigenetics, alternate splicing, gene expression levels, and protein-folding

Chapter 6 is based on a study that was published in the *Journal of Proteomics & Bioinformatics*.[1]

processes create a great deal of complexity. Qualitative trait analysis, which is the GWAS model most commonly reported in the literature, ignores these factors. As of mid-2015, more than 2,300 human GWAS have examined more than 210 diseases and traits and have reported more than 1,200 SNP associations.[5] Most of these GWAS employed a single-gene model that assumes that each locus acts independently of the others.

Many researchers believe that complex diseases involve multiple genes and their interactions.[6,7] Although GWAS have had some success in identifying genetic variants underlying complex diseases, most existing studies are based on limited single-locus approaches, which detect SNPs based on their marginal associations with a qualitative disease diagnosis.

Classical statistical tests derived from case-control experiments involving two loci that use a Pearson χ^2 test or logistic regression are commonly used as single-locus tests for GWAS and can be used in searching for pairwise interactions. Marchini and colleagues showed that explicitly modeling interactions between loci for GWAS with hundreds of thousands of markers is computationally feasible.[8] They also showed that simple methods that explicitly consider interactions can actually achieve reasonably high power with realistic sample sizes under different interaction models with some marginal effects, even after adjusting for multiple testing using the Bonferroni correction. However, the genotype-phenotype scenarios addressed in a study by Marchini et al. had substantially larger effects than those that we examine here.[8] Specifically, we focus on low-effect loci—those with low relative risk of association with disease diagnosis—because the evidence suggests they are common.[9] We also focus on theoretical examples of epistasis that are affected by the mode of inheritance without assuming an additive inheritance model.

An overarching goal of this study was to review the evidence of whether statistical methods based on single-gene models can effectively identify genotype-phenotype associations for multigene processes. Detecting such associations is particularly difficult for genetic variants with modest impacts on risk. Consequently, our experiments specifically investigated scenarios involving low-risk genetic variants and assessed whether multigene scenarios could be a source of the "missing heritability" observed using single-gene models.[10] We also examined the impact of two recent studies that collaborated in the development of novel tests for measuring interaction between two

linked (in epistasis) or unlinked loci.[11,12] These studies purport to have higher powers to detect interaction than classical logistic regression models.

Our investigations demonstrate that for low-effect loci, single-gene models of association fail to identify many associations because the interacting locus masks the effect on the index locus. For the scenarios we tested, our results also support assessments by Wu and colleagues and Ueki and colleagues that analytical methods that assume statistical interactions between loci are more powerful than single-loci models.

In this chapter, we will refer to markers as "loci," but more broadly, they could also be viewed as genes, SNPs, or haplotypes.

Epistasis Analysis

One way to extend the single-gene model to accommodate multiple genes involves studying gene pairs and their epistatic relationships. Epistasis analysis is the genetic methodology used to identify which genes act in a particular cellular process or pathway and to establish an order-of-function map that reflects the sequence in which those genes act. The analysis typically involves determining for a pair of genes whether the phenotype of a double mutant resembles that of a single mutant or whether it is a novel phenotype. Knowing what type of pathway is being investigated can help establish the type of relationship between the two genes.

Two types of pathways can be defined: substrate-dependent and switch-regulatory. Substrate-dependent pathways consist of a specific series of positive reactions, each of which involves some gene product (e.g., an enzyme) acting on a substrate produced in the previous step in the pathway and ultimately producing some final outcome. Switch-regulatory pathways consist of genes encoding negative or positive regulatory factors that alternate between "on" and "off" states depending upon upstream signaling events, thereby affecting some downstream response. Because substrate-dependent pathways comprise only positive factors whereas switch-regulatory pathways can comprise both positive and negative factors, interpreting results from epistatic studies is typically less complex for substrate-dependent pathways. Therefore, for the sake of simplicity, this analysis focuses on substrate-dependent pathways.

A number of studies argue that interacting loci may be the norm and not the exception. For example, Templeton and colleagues report that experience has revealed that most complex traits depend on more than one locus.[13] Their

study focuses on how often interactions among the loci play a significant role in the mapping from genotype to phenotype, given that the phenotype is influenced by two or more loci. They discuss a number of candidate scenarios, including coronary artery disease, in which the ApoE gene has been shown to affect males and females differently. Even the reported Mendelian trait sickle-cell anemia is commonly presented as a single nucleotide trait. A study by Gilbert-Diamond and colleagues indicates that gene-gene interactions (epistasis) are a significant complicating factor in the search for disease susceptibility genes.[14]

Objective

This chapter investigates epistatic interactions in a GWAS context using a qualitative association model. The purpose of this exercise is to determine the statistical methods and models that reliably predict associations between a qualitative phenotype (specifically, a disease diagnosis, coded as "case," for a positive diagnosis, or "control," for a negative diagnosis) and a pair of interacting genes. As with our other work, we use the concept of relative risk, the ratio of the probability of a positive diagnosis given a specific genotype and epistatic model (EM) divided by the probability with no risk present (i.e., P). The value of P is specified exogenously.

Methods

We employed a Monte Carlo–based simulation method to generate synthetic data corresponding to a variety of possible epistatic models for substrate-dependent pathways. The method takes into account factors known to influence association measurements in GWAS, including the relative risk of association, disease prevalence in nonrisk populations, inheritance properties of the simulated loci, and most important, the epistatic relationship of the simulated loci. We then analyzed the simulated gene data to assess the influence of these individual factors on statistical power in the context of GWAS. There were two advantages to using simulated data. First, the association-affecting factors were isolated and could be linked to the affecting locus. Second, we could choose any specific statistical method to perform the association assessment.

Epistatic Models of Inheritance

Table 6.1 defines four possible EMs for substrate-dependent pathways, as described in the literature.[15] Let *gene1* and *gene2* be distinct genes with varying genotypes that affect the production of a common gene product, P, ultimately influencing a phenotype (diagnosis of disease). Mutation of *gene1* results in a level of expression X of P and a relative risk Φ_a of exhibiting the disease phenotype. Similarly, mutation of *gene2* results in a level of expression Y of P and a relative risk Φ_b of exhibiting the disease phenotype. The phenotype of the *gene1gene2* double mutant varies according to the EM. If *gene1* acts upstream of *gene2* in the pathway leading to P (EM 1), the double mutant exhibits the phenotype of the *gene1* single mutant (*gene1* is epistatic to *gene2*). Conversely, if *gene2* acts upstream of *gene1* in the pathway leading to P (EM 2), the double mutant exhibits the phenotype of the *gene2* single mutant (*gene2* is epistatic to *gene1*). If *gene1* and *gene2* function in parallel pathways leading to P (EM 3), the double mutant exhibits a novel, more extreme level of P expression, Z, with associated relative risk Φ_{ab}. Finally, if *gene1* and *gene2* act at the same step in the pathway leading to P (EM 4), the observed phenotype can be either one of the phenotypes of the single mutants or a novel phenotype.

Table 6.1. Epistatic models for substrate-dependent pathways

		Phenotype of *gene1* single mutation	**Phenotype of *gene2* single mutation**	**Phenotype of *gene1gene2* double mutation**
Model 1	*gene1* → *gene2* →	X (Φ_a)	Y (Φ_b)	X (Φ_a)
Model 2	*gene2* → *gene1* →	X (Φ_a)	Y (Φ_b)	Y (Φ_b)
Model 3	*gene1* → *gene2* →	X (Φ_a)	Y (Φ_b)	Z (Φ_{ab})
Model 4	*gene1*, *gene2* →	X (Φ_a)	Y (Φ_b)	X (Φ_a), Y (Φ_b), or Z (Φ_{ab})

To simulate the EM scenarios in Table 6.1 in terms of the contributing locus genotypes, we referred to classical genetics material found in Klug and Cummings.[16] In each of the models, there are either two or three possible phenotypes. In our scenarios, there are only two phenotypes (a positive or negative diagnosis), but the risk of a diagnosis depends on the specific pairings of the genotypes. Table 6.2 outlines the expected risks associated with each possible combination of the wild-type (A and B) and mutant (a and b) alleles of *gene1* and *gene2* for EM1, taking into account the mode of inheritance acting at each locus. Because EM2 complements EM1, and EM3 and EM4 are subsumed by EM1, we limited our analysis to EM1. We used Table 6.2 to generate synthetic data sets representing the various scenarios and then examined our ability to link (associate) the phenotypes with the contributing genotypes.

Table 6.2. Epistatic model 1 depicted in terms of risk associated with various genotype combinations

	gene1	**D**	**D**	**R**	**R**
MOI	***gene2***	**D**	**R**	**D**	**R**
gene1	***gene2***	**Risk (¥)**	**Risk (¥)**	**Risk (¥)**	**Risk (¥)**
AA	BB	1	1	1	1
AA	Bb	Φ_b	1	Φ_b	1
AA	bB	Φ_b	1	Φ_b	1
AA	bb	Φ_b	Φ_b	Φ_b	Φ_b
Aa	BB	Φ_a	Φ_a	1	1
Aa	Bb	Φ_a	Φ_a	Φ_b	1
Aa	bB	Φ_a	Φ_a	Φ_b	1
Aa	bb	Φ_a	Φ_a	Φ_b	Φ_b
aA	BB	Φ_a	Φ_a	1	1
aA	Bb	Φ_a	Φ_a	Φ_b	1
aA	bB	Φ_a	Φ_a	Φ_b	1
aA	bb	Φ_a	Φ_a	Φ_b	Φ_b
aa	BB	Φ_a	Φ_a	Φ_a	Φ_a
aa	Bb	Φ_a	Φ_a	Φ_a	Φ_a
aa	bB	Φ_a	Φ_a	Φ_a	Φ_a
aa	bb	Φ_a	Φ_a	Φ_a	Φ_a

MOI = mode of inheritance.

Generation of Epistatic Synthetic SNP Data

The data generation method we used applies only to autosomal genes. Furthermore, because our simulation process assumed epistatic behaviors involving two interacting loci, we expect that the findings would apply to genes exhibiting these types of interactions. We began generating data by considering disease penetrance. We define P as the prevalence of a specific trait due to nongenetic factors. We designate a as the risk allele and A as the allele without risk for *gene1*. Similarly, we designate b as the risk allele and B as the allele without risk for *gene2*. Following the procedure of Iles,[17] we can then define the risk of disease as the ratio of the probability of a case given a and/or b divided by the probability of a case given no risk allele, which is P:

$$\Psi = \Pr(\text{case} / a, b) / P . \tag{6.1}$$

Generating the synthetic data set was straightforward, using the relationships between P and risk for the different epistatic categories. Initially, we assigned values to the following variables:

- n = the target number of cases and controls in a given experiment;
- P = the disease prevalence in subjects without genetic risk of a diagnosis;
- Φ_a , Φ_b = the relative risks (1.10, 1.15); and
- G = {g1, g2, g3}, a set of genotype distributions obtained from actual SNP data.[18]

Our general strategy was to randomly select a genotype and assign a relative risk (Φ_a, Φ_b) based on Table 6.2. Using the prevalence (P) assumption, we then assigned a case or control code (1, 0). A detailed description of the process follows:

1. Using the master genotype distribution G, draw at random a genotype (g1, g2 or g3) for *gene1*.
2. Repeat this process for *gene2*; that is, draw at random a genotype (g1, g2 or g3) for *gene2*.
3. Using Table 6.2, select the risk value ¥ of a case for the epistatic model being considered.
4. Based on ¥ and P, define the probability of a case to be

$$x = ¥ \times P . \tag{6.2}$$

5. Using the estimate of x from equation (6.2), assign a case (0) or control (1) designation at random. Note that using the 12 different EM/MOI combinations outlined in Table 6.2 for EMs 1–3, cases should be linked to both genetic loci, and this association should be identifiable via appropriate statistical procedures. Disease risk depends on specific and unknown disease mechanisms. A relative risk of 1.7 is considered strong and is associated with positive replication.[19] However, a risk of 1.3 is considered to be a realistic assumption for complex diseases.[20] However, many instances of risk < 1.1 are reported in the literature. We limited our focus to a relative risk range of 1.10 to 1.25 and were particularly interested in cases with low relative risk. Note that implicit in equation (6.2) is a definition of prevalence as the proportion of cases that are present where no genetic risk is assumed.
6. Continue the process until *n1* cases and *n2* controls have been generated (note that in this example *n1* = *n2*, but the procedure can be tailored to specific *n1*/*n2* targets).

Statistical Models

Using the assumptions presented in Table 6.2, we generated 1,000 replicates of genotypic and phenotypic data for each MOI pair for EM 1 using different sample sizes and risks. We then investigated the power of different statistical models to detect genotype-phenotype associations. We analyzed models that test each gene independently for association with the phenotype and models that test pairs of genes with and without interaction terms for association.

Single-Gene Methods: Cochran-Armitage Trend Test. The Cochran-Armitage (CA) trend test is often used as a genotype-based test for case-control genetic association studies, as described by Purcell and colleagues.[21] More generally, it is used in categorical data analysis to detect the presence of an association between a variable with two categories (e.g., a diagnosis) and a variable with k categories (e.g., a genotype). The CA trend test modifies the chi-square test to incorporate a suspected ordering in the effects of the k categories of the second variable. For example, one could order the number of mutated alleles as "zero," "one," and "two" and conjecture that the allele effect will not become smaller as the dose increases.

As described by Zheng and Gastwirth, the CA trend test has three flavors: dominant (CA-D), recessive (CA-R), and additive (CA-A).[22] Using the notation in Table 6.3 below to define the 2 × 3 table of case-control counts

stratified by genotype, a test statistic ($T^2(x)$) for the three variations of the CA trend methods can be defined as:

$$T^2(x) = \frac{n\,[\Sigma_{0,1,2}\,\{x_i\,(s\,r_i - r\,s_i)\}]^2}{[r\,s\,(\Sigma_{0,1,2}\,n\,\{x_i\,x_i\,n_i\} - \{\Sigma_{0,1,2}\,(x_i\,n_i)^2\})]}. \qquad (6.3)$$

The variables *ri*, *si* and *ni* in equation (6.3) are defined in Table 6.3. The variable x_i represents the specific test, namely $x_0 = 0$, $x_2 = 1$ and $x_1 = .5$.

Table 6.3. Terms defined in equation (6.2)

	AA	Aa	aa	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	*N*

AA = major genotype; **Aa** = heterozygote genotype; **aa** = minor genotype.

Under the null hypothesis of no association, $T^2(x)$ has an asymptomatic χ^2 distribution with 1 degree of freedom. We applied the above test to both *gene1* and *gene2*.

Two-Gene Models: Pearson Test. The two-gene, case-control test is derived from the classical case-control test of epidemiology described by Jewell.[23] As with all of the tests, this test compares subjects who have a condition (the "cases") with subjects who do not have the condition but are otherwise similar (the "controls"). As in the CA test described previously, the Pearson χ^2 test is used in categorical data analysis when testing for the presence of an association between a variable with two categories (e.g., a positive or negative diagnosis) and two variables with *k* categories (e.g., three genotypes). For this test, the columns are the nine combinations of genotypes and the rows are the two case-control designations. The central idea is to compute the theoretical frequencies for all 18 cells from the marginal totals and then test for statistically significant differences between the theoretical and observed frequencies. This test also uses a χ^2 test with $(nr - 1) \times (nc - 1) = 8$ degrees of freedom.

Two-Gene Models: The Method of Wu et al., as Refined by Ueki et al. Wu et al. developed two novel statistics, refined by Ueki et al., designed to test interactions between linked or unlinked loci without including the influence of main effects.[11,12]

The two-locus linked test, T_{IH} linked, is defined as

$$T_{IH} = \left(\hat{I}_{GH}^{H}\right)^{2} / V\left(\hat{I}_{GH}^{H}\right),$$

where

$$\left(\hat{I}_{GH}^{H}\right) = \left[M1 - M2\right]^{2}$$

$$M1 = \log \frac{\hat{P}_{11}^{A}\hat{P}_{22}^{A}}{\hat{P}_{12}^{A}\hat{P}_{21}^{A}}$$

$$M2 = \log \frac{\hat{P}_{11}^{N}\hat{P}_{22}^{N}}{\hat{P}_{12}^{N}\hat{P}_{21}^{N}}$$

$$V\left(\hat{I}_{GH}^{H}\right) = V1 + V2$$

$$V1 = \frac{1}{4n_{A}}\left[\frac{1}{P_{11}^{A}} + \frac{1}{P_{12}^{A}} + \frac{1}{P_{21}^{A}} + \frac{1}{P_{22}^{A}}\right]$$

$$V2 = \frac{1}{4n_{G}}\left[\frac{1}{P_{11}^{N}} + \frac{1}{P_{12}^{N}} + \frac{1}{P_{21}^{N}} + \frac{1}{P_{22}^{N}}\right].$$

The second test, T_{IH} unlinked, assumes that the two loci are unlinked and is defined as

$$T_{IH} = \left(\hat{I}_{GH}^{H}\right)^{2} / V\left(\hat{I}_{GH}^{H}\right),$$

where

$$\left(\hat{I}_{GH}^{H}\right) = \left[M1\right]^{2}$$

$$V\left(\hat{I}_{GH}^{H}\right)^{2} = V1 \quad .$$

Results

This study investigated the effect that polygene interactions have on association predictions in a GWAS context. We used statistical models that appear in the literature to generate predictions. Some of the models were single-gene, inheritance-specific models; that is, they assumed that a single additive or recessive or dominant gene produced the diagnosis. Other models were inheritance agnostic and assumed that a pair of interacting genes produced the diagnosis. To implement this investigation, we fixed the risk of the upstream gene of EM1, *gene1*, to a low but detectable 1.10 risk level. Simultaneously, we varied the risk on the downstream gene, *gene2*, from 1.00 (no risk) to 1.20, a level that is twice as high as the risk of *gene1*. Note that a no-risk gene is inconsistent with the purpose of Table 6.2, which identifies the interactions

between two genes; however, we use this scenario to describe an endpoint in our assessment.

Table 6.4 presents a power analysis for *gene1* of the simulated EM1 data using six different statistical tests when the downstream gene has no risk of disease (the single-gene scenario). The first three columns correspond to three different versions of the single-gene CA test with different inheritance assumptions: additive (CA-A), dominant (CA-D), or recessive (CA-R). Each test was applied to both the upstream and the downstream gene. The last three columns of Table 6.4 present the results for the two-gene tests.

Table 6.4. Model comparisons for EM 1, *N* = 12,500, P=.4, $\Phi_a = 1.10$, $\Phi_b = 1.00$

MOIs/stat model	CA-A *gene1*	CA-D *gene1*	CA-R *gene1*	CC	T_{IH} linked	T_{IH} unlinked
D-D	59.34	**73.63**	0.00	50.40	0.00	0.30
D-R	59.35	**73.40**	0.05	51.70	0.00	0.30
R-D	9.05	0.00	**23.70**	9.05	0.00	0.20
R-R	7.65	0.00	**23.30**	9.20	0.05	**27.15**

CA-A = Cochran-Armitage additive test; CA-D = Cochran-Armitage dominant test; CA-R = Cochran-Armitage recessive test; CC = case-control association test ; MOI = mode of inheritance.

Table 6.4 indicates the following results:

- Single-gene tests work better (from a statistical power perspective) than two-gene tests for single-gene scenarios (i.e., low risk of disease for *gene1*, no risk for *gene2*), because the additional degrees of freedom used by the two-gene test provide no benefit when there is no additional risk of disease from the second interacting gene.
- In general, the MOI of the upstream gene determines which test is optimal (optimal values are bolded in red) with the dominant version of the CA test being optimal for dominant genes and the recessive version of the CA test being optimal for recessive genes. Accordingly, the commonly used additive CA test (CA-A) is never optimal unless the MOI of the gene is additive.[3]
- Unexpectedly, when both genes are recessive, the unlinked refined test is optimal,[12] although the risk of the second locus is null.

In contrast, Table 6.5 presents the results for the case in which the risk from the downstream gene is twice the risk of the upstream gene.

Table 6.5. Model comparisons for EM 1, *N* = 12,500, P=.4, Φ_a = 1.10, Φ_b = 1.20

MOIs/stat model	CA-A *gene1*	CA-D *gene1*	CA-R *gene1*	CC	T_{IH} unlinked	T_{IH} linked
D-D	0.20	0.45	0.00	**80.87**	**96.56**	**93.31**
D-R	25.45	38.15	0.00	**68.65**	**63.05**	**53.35**
R-D	0.00	0.00	0.00	**99.95**	**61.90**	**49.50**
R-R	1.65	0.05	8.35	**81.75**	**27.30**	**27.25**

CA-A = Cochran-Armitage additive test; CA-D = Cochran-Armitage dominant test; CA-R = Cochran-Armitage recessive test; CC == case-control association test ; MOI = mode of inheritance.

Table 6.5 indicates the following results:

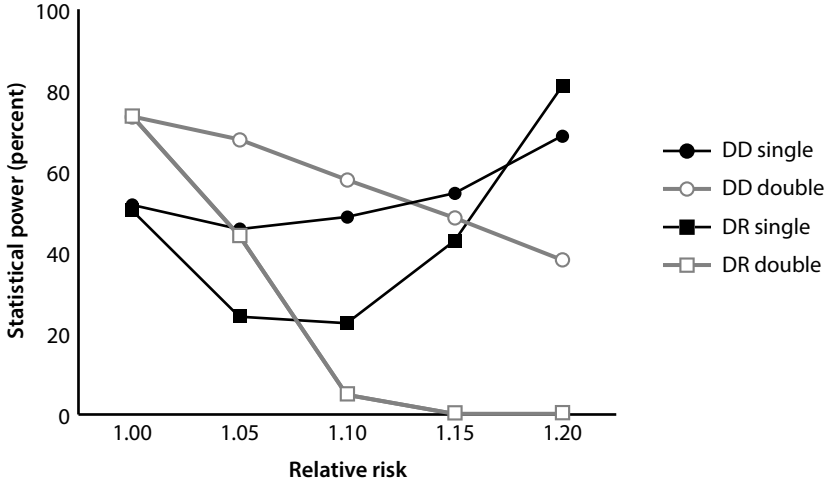
- The Pearson two-gene test is optimal for all MOI submodels of the EM1 model, except when both genes are dominant. In this case, the unlinked refined Wu et al. test is optimal.
- The risk conveyed by *gene2* has apparently masked the contribution of *gene1* and the power to predict an association between *gene1* and diagnosis using single-gene models is very low, below 3 percent in all cases except submodel D-R, where it is below 30 percent. This finding suggests that *gene1* is unlikely to be associated with a diagnosis using single-gene models.

Figure 6.1 provides estimates of the statistical power (*y*-axis) to predict association between *gene1* and diagnosis given different risk values for *gene2* (*x*-axis) for the dominant submodels (D-D and D-R) for EM1. The results presented in Figure 6.1 correspond to the best single-locus and two-locus tests. Note that the risk value for *gene1* is fixed (1.10). Figure 6.2 presents the same information for the recessive submodels (R-D and R-R). Both Figures 6.1 and 6.2 identify the crossover risk, which is the risk value at which the single-gene (optimal model) and the two-gene model have the same power.

These figures suggest that beyond a risk value of 1.05–1.12 (depending on the MOI), single-gene tests are no longer as effective (from a power perspective) as two-gene tests. Furthermore, the power of two-gene tests improves as the risk of the downstream gene increases, whereas the power of single-gene tests progressively declines as the risk from *gene2* increases.

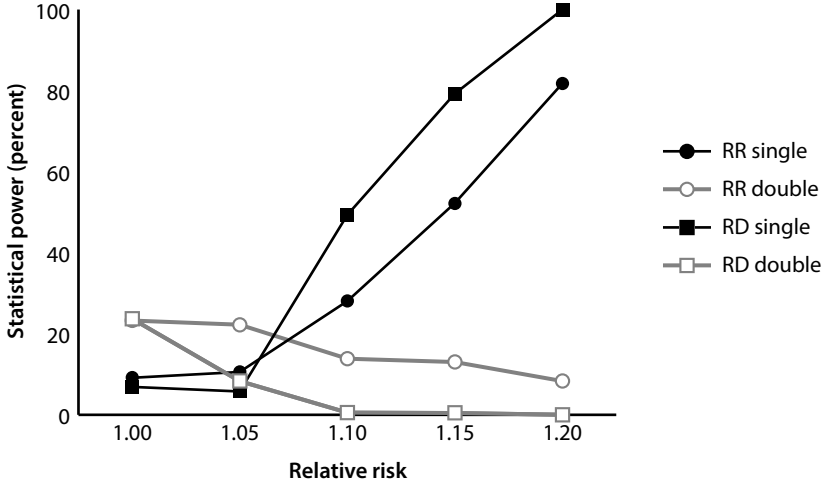
We repeated the same analysis for EM 1 with the downstream gene (*gene2*) risk fixed at 1.1. This time we varied the risk levels of *gene1* and applied the single-gene tests to *gene2*. Table 6.6 presents the results when the risk from the upstream gene is null. Again we acknowledge that a no-risk gene is inconsistent with the purpose of Table 6.2, but this scenario describes an endpoint.

Figure 6.1. Dominant *gene1* analysis: D-D crossover risk = 1.07, D-R crossover risk = 1.12



DD single = *gene1* dominant, *gene2* dominant, single-gene test; DD double = *gene1* dominant, *gene2* dominant, two-gene test; DR single = *gene1* dominant, *gene2* recessive, single-gene test; DR double = *gene1* dominant, *gene2* recessive, two-gene test

Figure 6.2. Recessive *gene1* analysis: R-D crossover risk = 1.05, R-R crossover risk = 1.07



RD single = *gene1* recessive, *gene2* dominant, single-gene test; RD double = *gene1* recessive, *gene2* dominant, two-gene test; RR single = *gene1* recessive, *gene2* recessive, single-gene test; RR double = *gene1* recessive, *gene2* recessive, two-gene test

Surprisingly, the results indicate that single-gene tests do not universally perform better (in a power sense) than the two-gene tests, even when there is no risk of diagnosis from *gene1*. Specifically, if *gene1* is recessive, the single-gene tests do as well as or better than the 8*df* χ^2 two-gene tests, but if *gene1* is dominant, the two-gene unlinked refined Wu et al.[12] outperforms the single-gene tests.

Table 6.6. Model comparisons for EM 1, *N* = 12,500, P=.4, $\Phi_a = 1.00$, $\Phi_b = 1.10$

MOIs/ stat model	CA-A *gene2*	CA-D *gene2*	CA-R *gene2*	CC	T_{IH} unlinked	T_{IH} linked
D D	1.85	3.55	0.10	32.12	**37.57**	26.75
D R	0.10	0.00	0.65	1.15	**11.10**	8.45
R D	41.20	**56.15**	0.00	52.60	11.65	8.05
R R	4.20	0.00	14.65	8.10	3.95	**23.90**

CA-A = Cochran-Armitage additive test; CA-D = Cochran-Armitage dominant test; CA-R = Cochran-Armitage recessive test; CC = ;case-control association test; MOI = mode of inheritance.

Table 6.7 presents the results for the case in which the risk from the upstream gene is twice the risk of the downstream gene and demonstrates that the two-locus, case-control test outperforms all single-gene tests and both of the refined Wu et al.[12] tests in this scenario.

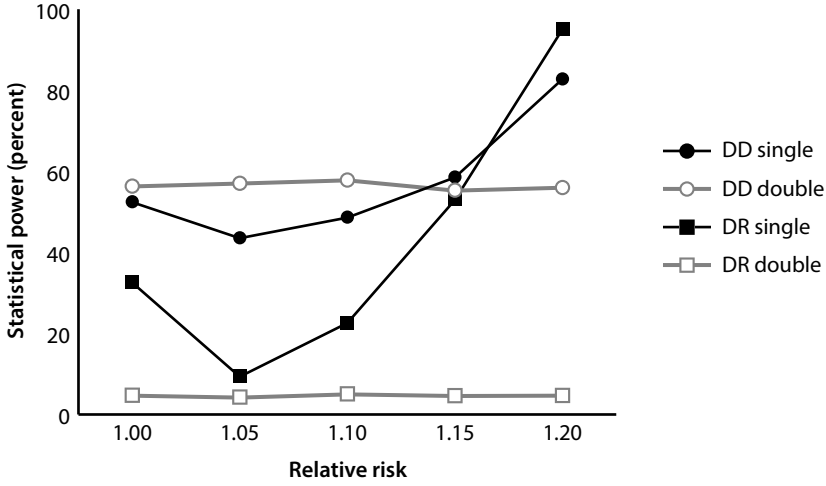
Table 6.7. Model comparisons for EM 1, *N* = 12,500, P=.4, $\Phi_a = 1.20$, $\Phi_b = 1.10$

MOIs/ stat model	CA-A *gene2*	CA-D *gene2*	CA-R *gene2*	CC	T_{IH} unlinked	T_{IH} linked
D D	2.00	4.40	0.10	**95.36**	38.22	26.70
D R	5.00	0.00	0.55	**100.0**	11.15	8.05
R D	42.90	57.80	0.00	**83.50**	11.45	9.50
R R	3.70	0.05	14.00	**83.80**	3.60	25.90

CA-A = Cochran-Armitage additive test; CA-D = Cochran-Armitage dominant test; CA-R = Cochran-Armitage recessive test; CC = = case-control association test ; MOI = mode of inheritance.

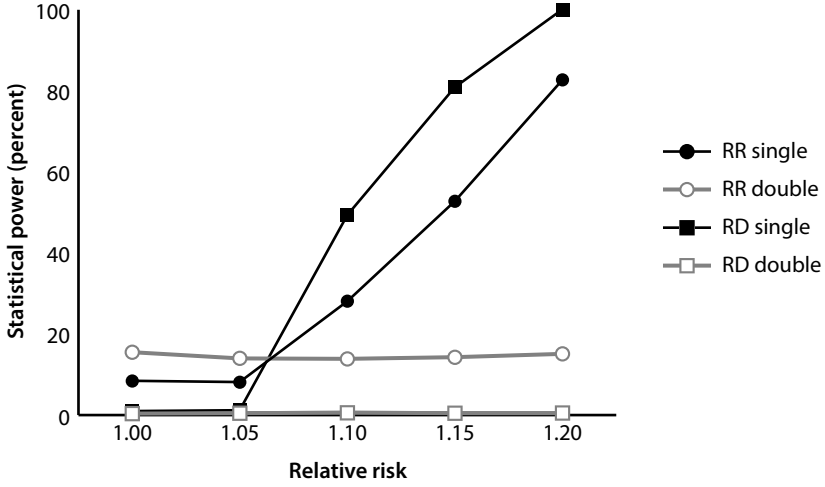
Figure 6.3 provides estimates of the power (*y*-axis) to detect association between *gene2* and disease diagnosis given different risk values for *gene1* [1.0 (no risk) to 1.20 (double the risk of *gene1*)]. Note that the risk of *gene2* is fixed at 1.10. Figure 6.3 presents the results for the dominant *gene2* submodels (D-D and R-D). Figure 6.4 presents the results for the recessive *gene2* submodels (D-R and R-R).

Figure 6.3. Dominant *gene2* analysis: D-D crossover risk = 1.0, R-D crossover risk = 1.12



DD single = *gene1* dominant, *gene2* dominant, single-gene test; DD double = *gene1* dominant, *gene2* dominant, two-gene test; RD single = *gene1* recessive, *gene2* dominant, single-gene test; RD double = *gene1* recessive, *gene2* dominant, two-gene test

Figure 6.4. Recessive *gene2* analysis: R-D crossover risk = 1.05, R-R crossover risk = 1.07



DR single = *gene1* dominant, *gene2* recessive, single-gene test; DR double = *gene1* dominant, *gene2* recessive, two-gene test; RR single = *gene1* recessive, *gene2* recessive, single-gene test; RR double = *gene1* recessive, *gene2* recessive, two-gene test

Figure 6.3 and 6.4 are consistent with the results from Figures 6.1 and 6.2 and further suggest that beyond risk value = 1.05, single-gene tests are no longer as effective from a power perspective as two-gene tests. Furthermore, the power of two-gene tests improves as the risk of the downstream gene increases. This is exactly the opposite of the scenario for single-gene tests, which decline in power as the risk of the downstream gene increases.

Discussion

Our investigation of epistatic scenarios involving low-risk loci indicates that for a given locus, single-locus tests are not as effective as two-locus tests for predicting associations if the risk value for a second interacting locus exceeds 1.05–1.12 (the crossover risk value varies depending on the genetic inheritance properties of the pair of loci). In general, the power of two-locus tests to detect associations improves as the risk value of the second locus increases, whereas the power of single-locus tests progressively declines. Disturbingly, for certain inheritance models and risk values, a true association between a locus and phenotype can be entirely masked by a second interacting locus when using single-locus tests. These finding are not unexpected and are consistent with previous findings reported by others.[6,24,25] However, single-gene models continue to be used as the core methods for detecting associations in a GWAS context. Our study is significant in that it provides a more exact estimate of the risk scenarios in which single-locus models are inferior.

Comparing the performance of the three different two-locus tests evaluated in this study, in most cases for EM1, the two-locus, case-control Pearson test is optimal. In certain scenarios (i.e., when both genes have a dominant MOI), the unlinked Wu et al. test (in which cases and controls are included) as refined by Ueki et al. is optimal.[11,12] This finding is somewhat surprising given that the modified Wu et al. test measures interaction effects exclusively, whereas the two-locus, case-control test includes main effects for both loci as well as interaction effects.

Despite the widespread recognition that single-locus tests are likely to be inferior to multilocus tests for GWAS of many diseases and phenotypes, an unresolved issue is how to construct a computationally practical test that takes into account interactions and enhances the detection of associations between a specific locus and the phenotype of interest. Wang and colleagues conducted an empirical comparison of five epistatic interaction detection methods, including a number of two-pass methods.[26] They indicate that each of the five

methods demonstrates unique utilities, but no single method is optimal, being simultaneously the most powerful and the most scalable and having the lowest type-1 error rate in every setting. When users want powerful results and are not concerned with computation cost, Wang and colleagues cite Zhang and colleagues' TEAM method as the best-performing algorithm.[26,27] However, researchers should note that even when limiting the number of interacting genes to two, $n \times (n - 1) / 2$ association calculations are required. For n = 500,000–1,000,000, the computational requirements of such an analysis are daunting but readily parallelizable.

Chapter References

1. Cooley P, Gaddis N, Folsom R, et al. Conducting genome-wide association studies: epistasis scenarios. J Proteomics Bioinform. 2012;5(10):245-251.
2. Kuo CL, Feingold E. What's the best statistic for a simple test of genetic association in a case-control study? Genet Epidemiol. 2010;34(3):246-253.
3. Cooley P, Clark R, Folsom R, et al. Genetic inheritance and genome wide association statistical test performance. J Proteomics Bioinform. 2010;3(12):321-325.
4. Manolio TA. Genomewide association studies and assessment of the risk of disease. *N* Engl J Med. 2010;363(2):166-76.
5. Burdett T, Hall P, Hasting E, et al. The NHGRI-EBI Catalog of published genome-wide association studies. 2015 [cited 2015 Nov 2]; Available from: www.ebi.ac.uk/gwas
6. Li J, Horstman B, Chen Y. Detecting epistatic effects in association studies at a genomic level based on an ensemble approach. Bioinformatics. 2011;27(13):i222-9.
7. Carlson CS, Eberle MA, Kruglyak L, et al. Mapping complex disease loci in whole-genome association studies. Nature. 2004;429(6990):446-52.
8. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet. 2005;37(4):413-7.
9. Suhre K, Shin SY, Petersen AK, et al. Human metabolic individuality in biomedical and pharmaceutical research. Nature. 2011;477(7362):54-60.
10. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.

11. Ueki M, Cordell HJ. Improved statistics for genome-wide interaction analysis. PLoS Genet. 2012;8(4):e1002625.

12. Wu X, Dong H, Luo L, et al. A novel statistic for genome-wide interaction analysis. PLoS Genet. 2010;6(9):e1001131.

13. Templeton AR. Epistasis and complex traits. In: Wolf JB, Brodie EDI, Wade MJ, editors. Epistasis and the evolutionary process. New York, NY: Oxford University Press; 2000. p. 41-57.

14. Gilbert-Diamond D, Moore JH. Analysis of gene-gene interactions. Curr Protoc Hum Genet. 2011;Chapter 1:Unit1 14.

15. Michels CA. Genetic techniques for biological research: a case study approach. Chichester, United Kingdom: John Wiley & Sons; 2002.

16. Klug W, Cummings M. Concepts of genetics. 7th ed. New York, NY: Prentice Hall; 2010.

17. Iles MM. Effect of mode of inheritance when calculating the power of a transmission/disequilibrium test study. Hum Hered. 2002;53(3):153-7.

18. Schymick JC, Scholz SW, Fung HC, et al. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol. 2007;6(4):322-8.

19. Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007;445(7130):881-5.

20. Ziegler A, Konig IR, Thompson JR. Biostatistical aspects of genome-wide association studies. Biom J. 2008;50(1):8-28.

21. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75.

22. Zheng G, Gastwirth JL. On estimation of the variance in Cochran-Armitage trend tests for genetic association using case-control studies. Stat Med. 2006;25(18):3150-9.

23. Jewell NP. Statistics for epidemiology. Boca Raton, FL: Chapman & Hall/CRC; 2004.

24. Hoh J, Wille A, Zee R, et al. Selecting SNPs in two-stage analysis of disease association data: a model-free approach. Ann Hum Genet. 2000;64(Pt 5):413-7.

25. Culverhouse R, Suarez BK, Lin J, et al. A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet. 2002;70(2):461-71.

26. Wang Y, Liu G, Feng M, et al. An empirical comparison of several recent epistatic interaction detection methods. Bioinformatics. 2011;27(21):2936-43.

27. Zhang X, Huang S, Zou F, et al. TEAM: efficient two-locus epistasis tests in human genome-wide association study. Bioinformatics. 2010;26(12):i217-27.

CHAPTER 7

Assessing Gene-Environment Interactions in Genome-Wide Association Studies (GWAS): Statistical Approaches

Philip Chester Cooley, Robert F. Clark, and Ralph E. Folsom

Overview

In this chapter, we address a scenario that uses synthetic genotype case-control data that are influenced by environmental factors in the context of a genome-wide association studies (GWAS). The precise way the environmental influence contributes to a given phenotype is typically unknown. Therefore, our study evaluates how to approach a GWAS that may have an environmental component. Specifically, we assess different statistical models in the context of a GWAS to make association predictions when the form of the environmental influence is questionable. We used a simulation approach to generate synthetic data corresponding to a variety of possible environmental-genetic models, including a "main effects only" model as well as a "main effects with interactions" model. Our method takes into account the strength of the association between phenotype and both genotype and environmental factors, but we focus on low-risk genetic and environmental risks that necessitate using large sample sizes (N = 10,000 and 200,000) to predict associations with high levels of confidence. We also simulated different Mendelian gene models, and we analyzed how the collection of factors influences statistical power in the context of a GWAS. Using simulated data provides a "truth set" of known outcomes such that the association-affecting factors can be unambiguously determined. We also test different statistical methods to determine their performance properties. Our results suggest that the chances of predicting an association in a GWAS is reduced if an environmental effect is present and the statistical model does not adjust for that effect. This is especially true if the

Chapter 7 is based on a study that was published in the RTI Press.[1] Minor text edits were made to this chapter. The analysis of the gene data has not changed.

environmental effect and genetic marker do not have an interaction effect. The functional form of the statistical model also matters. The more accurately the form of the environmental influence is portrayed by the statistical model, the more accurate the prediction will be. Finally, even with very large samples sizes, association predictions involving recessive markers with low risk can be poor.

Introduction

In recent years, scientists and researchers have increasingly used GWAS to unravel the genetic factors that influence important phenotypes such as disease presence and predisposition. The hypothesis GWAS follows is that if genetic variations are more frequent in people with a given disease, the variations are likely *associated* with the disease. In general, GWAS apply univariate statistical tests to each gene marker or single nucleotide polymorphism (SNP) as an initial step. This SNP-based test is statistically straightforward, and the core tests for assessing the associations are standard methods (e.g., χ^2 tests, regression) that have been studied outside of and within the GWAS context. Kuo & Feingold describe the most commonly used statistical methods applied to GWAS.[2] All the tests they cite are single-locus tests.

The popularity of the GWAS approach is testimony to its simplicity; however, it obscures the important issue of whether a single-gene model is conducive to unraveling the workings of the biosynthetic pathways of a phenotype. In the preceding chapter, we demonstrate that if two genes are in epistasis the likelihood of identifying the weaker (in terms of risk) of the two is diminished if a single-gene model is used in this context. We would anticipate a similar finding involving genes and environmental interactions namely that the risk weaker effect would be dominated by the stronger effect.

Researchers can use classical statistical tests derived from case-control experiments to determine whether two loci associate in a GWAS context. Both a Pearson χ^2 test and tests involving logistic regression can be used to examine for pair-wise interaction assumptions.

In this study, we focused on low-effect loci with low relative risks of association with disease diagnosis, because the evidence suggests these are common.[3] Most GWAS report only small changes in disease risk (1.1 to 1.5). It has also been reported that relative risks underestimate the true risk and the corresponding effect size.[4]

Note that this chapter does not assess the multilocus scenario, which is the focus of Chapter 8. Nor does it account for the scenario involving multiple

loci that are tested simultaneously. This scenario, also discussed in Chapter 8, requires an adjustment to the *p*-value threshold via a Bonferroni correction.[5] This simple procedure (dividing the *p*-value threshold by the number of test) assumes statistical independence between tests which is not true. Hence the correction represents an "overcorrection," leading to higher-than-necessary type II error rates. We note that the correction does not affect our type I or type II assessment because all of our examples have been generated with positive genetic associations (even if some of these associations are very difficult to detect). Thus, all scenarios are associated, and the issue is whether that association can be detected and the statistical procedures that perform most effectively.

The word "risk" can have a variety of meanings. In an environmental context, it means a hazard based on an exposure to a chemical or pollutant such as tobacco smoke. In another context, risk is interpreted more narrowly to mean the probability of an adverse consequence (e.g., an adverse event such as a disease). The term "environmental risk" in this study is used broadly; we define it as any process that contributes to a disease diagnosis that is not genetic in origin. For example, environmental risks can represent exposure to chemicals or pollutants—or a subject's age.

Our overarching goal was to identify which statistical methods best identify genotype-phenotype associations when environmental effects also influence the association. Detecting such associations is particularly difficult for genetic variants with modest impacts on risk. Consequently, our experiments specifically investigated scenarios involving low-risk genetic variants and assessed whether environmental influences with varied levels of risk could be a source of the "missing heritability" observed using single-gene models.[6] Not surprisingly, our investigations demonstrated that the best statistical method (with respect to statistical power) depends on whether there are interactions between the genotype and environmental factors and how well the specified statistical model matches the environmental effect associated with the phenotype. In summary, the simulated data set provides a truth set for assessing the sensitivity of the effect of the statistical method and the predicted association. Establishing the genotype-to-phenotype connections without using a simulation approach is difficult to impossible. Although our study results demonstrate a number of obvious "truths," a number of unexpected results may lead researchers to more powerful statistical approaches that can establish the validity of the simulation approach.

Background

Many complex diseases (e.g., diabetes, asthma, cancer) are affected in part by interactions between genes and environmental factors. However, investigators conducting GWAS typically do not investigate the influence of environmental factors as part of the GWAS process.

There have been several notable exceptions. For example, a study by Terry and colleagues showed a significant interaction between smoking status and the specific gene for lung cancer.[7] Another study, by Stern and colleagues found smoking status to be an effect modifier of the association between a codon and the risk of bladder cancer.[8] Understanding the relationship between genetic polymorphisms and environmental exposures can greatly aid investigators in detecting high-risk subgroups in the population and provide better insight into pathway mechanisms for complex diseases.

Current GWAS methods are designed to detect main effects, that is, direct associations of an SNP or clusters of SNPs with disease.[9,10] In the context of complex diseases, examining main effects only could miss important genetic variants specific to subgroups of the population.

Gene-Gene Interaction Studies

Lichtenstein and colleagues studied twins and sought to connect hereditary factors to the causes of sporadic cancer.[11] The study assessed the risks of cancer at 28 anatomical sites for twin children of a parent who has cancer. Statistical modeling was used to estimate the relative importance of heritable and environmental factors in causing cancer at 11 of those sites. A major finding was that inherited genetic factors make a minor contribution to susceptibility for most types of neoplasms, indicating that the environment plays the principal role in causing sporadic cancer. The relatively large effect of heritability in cancer at a few sites (such as prostate and colorectal cancer) suggests major gaps in our knowledge of the genetics of cancer.

Another large study, by Pearce and colleagues that also focused on cancer attempted to link several well-established environmental risk factors for ovarian cancer and the results of a recent GWAS that identified six variants that influence disease risk.[12] They pooled data from 14 ovarian cancer case-control studies, and then conducted stratified analyses of each environmental risk factor to evaluate the presence of interactions for all histological subtypes. They fit a multivariate model to examine the association between all environmental risk factors and genetic risk score on ovarian cancer risk.

The results indicated no strong statistical evidence of interaction between the six SNPs or genetic risk score and the environmental risk factors on ovarian cancer risk.

A large bladder cancer study reported by Rothman and colleagues demonstrated interactions due to smoking using a logistic regression (LR) adjusted for age.[13] This study coded the genotype variable as a count of minor alleles conforming to our Models 1, 2, and 3 described here. Lindstrom and colleagues, in a study involving prostate cancer, found no contribution from a number of environmental factors.[14] This study used a number of LR models similar to those we used in our analysis. Another study by Yu and colleagues developed a Bayesian framework to investigate the influence of multiple loci simultaneously on disease risk.[15] Yu and colleagues' "full" model consisted of a standard LR model that treats the genotype variable as a categorical variable and specifies a main effect with interactions model.

Patel and colleagues used GWAS to examine type 2 diabetes, a second disease with a strong interplay between environmental and genetic factors.[16] Genetic loci discovered through GWAS in this and other studies explained only a small portion of the disease risk variance; some of the unexplained risk was likely due to gene-environment interactions. The study suggested that the adverse effect of several type 2 diabetes loci may be abolished or at least attenuated by higher physical activity levels or healthy lifestyle, whereas low physical activity and the typical Western diet may augment it. This study used data from two surveys from the Centers for Disease Control and Prevention's National Health and Nutrition Examination Survey (NHANES). They used a GWAS to screen 18 genetic loci and type 2 diabetes for statistical interactions that were associated with disease. They describe their investigation as an environment-wide association study (EWAS), and they used data sets from four cohorts from the NHANES. Because the four cohorts were analyzed individually, the number of environmental factors varied among them.

The models used by Patel and colleagues were logistic regression examples that were adjusted for age, sex, body mass index, and race.[16] The results identified eight potential disease gene–environmental factor interactions. One interaction (trans-β-carotene) was particularly significant. The per–risk-allele effect sizes, after adjusting for age, sex, body mass index, and race for subjects with low trans-β-carotene levels, were 40 percent greater than the marginal genetic effect size of the SNP. They also found a strong interaction between an

SNP and a nutrient found in corn oil, which conveyed a 20 percent higher risk than the SNP alone did.

A study by Murcray and colleagues performed a general methodological study that focused on identifying SNPs that demonstrate heterogeneity between subgroups defined by some environmental exposure.[17] They describe a two-step approach for detecting loci involved in gene-environment interactions that is performed independently of any initial scans for main effects. They expanded on the traditional test for gene-environment interaction in a case-control study by incorporating a preliminary screening step constructed to efficiently use all available information in the data. They claim that their two-step approach is more powerful than the standard test of interaction across a wide range of models and consequently is more robust to changes in environmental exposure and minor allele frequency than the traditional one-step test for identifying highly significant SNPs. The difficulty with most methods, including theirs, is that it is not a "data mining" method. The specific environmental factor and the form of that factor have to be established prior to analysis. This has proven to be a difficulty with our methods as well. The specific environmental factor or factors to include in the model greatly affect the power of the tests. Specifically, researchers should use some combination of the literature and/or data mining activities to establish the form of the environmental effect model (step function or linear) on the logistic scale.

A study by Cornelis and colleagues provides a comparative study of several logistic regression–based tests of gene-environment (G-E) and G × E interactions.[18] All seven methods compared in their paper assumed a log-additive mode-of-inheritance model for each SNP. This differs from our methods, in which the mode of inheritance was agnostic. Cornelius et al. do not identify a preference for any of the seven methods and instead indicate that preference would depend on the goal of the study. They also explored methods investigating environment effects only in subjects with a positive phenotype case (i.e., case-only studies).

Finally, Kraft and colleagues performed a study, which was similar in content to the Cornelius and colleagues study in that it also focused on log-additive gene models, that formulated a likelihood ratio test of association between disease and locus with the possibility that the genetic effect may be modified by an environmental factor.[19] The specific environment model they

investigated was similar to one of the experiments examined in our study—namely, a chemical spill—because it was an all-or-nothing type of exposure.

In summary, all of the methods cited previously use logistic regression models. This is a result of the flexibility of that approach with respect to treating multiple genetic, environmental, and interaction variables as simultaneous effects. Also, all of the studies attempted to characterize how the environmental effects influenced the association outcome. In general, the effects could be characterized as single environmental exposures that triggered the risk of association immediately, or as a single risk that accumulated as the subject aged. Because each of these two types of risks could depend (or not) on a genetic predisposition, we decided to investigate these four categories of risk as separate possibilities. Because our study uses synthetic gene-environmental data with specified (known) risk, we are able to characterize data mining strategies in terms of their statistical power.

Methods

Overview

We simulated genetic and environmental interactions in a GWAS context using a qualitative association framework to determine which statistical methods and models reliably predict associations between a qualitative phenotype (specifically, a disease diagnosis, coded as "case" for a positive diagnosis or "control" for a negative diagnosis) and a gene paired with an environmental influence. As with our previous work, the concept of relative risk is the basis for this investigation.[20] We define the *genetic relative risk* (Φ) of a wild-type genotype to be the ratio of the probability of a positive diagnosis given an occurrence of a (wild-type) genotype divided by the probability of disease in the absence of the disease genotype. We also define the *environmental risk* (Π) as the ratio of the probability of a positive diagnosis given an exposure divided by the probability of a positive diagnosis in unexposed subjects. The values of Φ and Π are specified exogenously and vary from low-risk to not-so-low-risk.

We generated 1,000 replicates of simulation data that depended on the two risk values (Φ and Π) for each of three gene models using a standard Bernoulli process and analyzed them in terms of the observed power profiles for a low alpha error ($\alpha \leq 10^{-8}$). The distribution of the number of alleles per genotype was randomized across replicates and was based on real data from the study by Schymick and colleagues used in previous chapters.[21] We biased the risk levels

to the low end of the risk continuum because these are more difficult scenarios and are typical of what has been observed in the literature.[3] To support these low risk levels, we fixed our sample size to $N = 10{,}000$ (5,000 cases and 5,000 controls) and $N = 200{,}000$ (100,000 cases and 100,000 controls) to determine whether it is possible to measure associations in low-risk, recessive inheritance scenarios. Other studies have used smaller values ($N = 6{,}000$) for comparable investigations.[13]

Generating the Synthetic SNP Data

We derived our data generation method from a study by Iles and Mendelian concepts of inheritance.[22] We specifically incorporated autosomal dominant, recessive, and additive inheritance patterns into the data. These data also depend on factors known to influence association measurements in the context of GWAS. Our simulation process assumes Mendelian-type inheritance patterns.

"Penetrance" was defined as the proportion of individuals without the risk allele who have a definable trait (phenotype). In other words, penetrance was a genotype-specific probability of being affected with the trait. We designated **a** as the risk allele and **A** as the allele without risk. Generating the synthetic data set using the relationships between penetrance and risk for different mode of inheritance (MOI) categories was straightforward; Chapter 3 and Cooley and colleagues provide additional detail.[23]

Initially, we supply as input data the following variables:

- n = the target number of cases and controls in a given experiment,
- P = the disease penetrance,
- Φ = the genotype relative risk (1.10, 1.15, 1.20),
- Π = the environmental relative risk, and
- the distribution of genotypes, which were drawn at random from a master set of genotype distributions obtained from real SNP data.[23]

In screening samples from the master set of genotype distributions, Chan and colleagues recommend that a minor allele frequency (MAF) threshold not be applied as a filter.[24] They argue that filtering MAFs out of the process because of low frequencies or to maintain Hardy–Weinberg equilibrium deviation has little effect on the overall false positive rate and, in some cases, filtering MAFs excludes SNPs. The effect of this step is to select a specific genotype distribution at random from the master distribution.

From the selected relative risk (Φ), penetrance (P), and MOI assumptions, we used the formulas in Table 7.1 to assign a case (1) or control code (0). This step converts the relative risk ratio (Φ) into the probability of a case (disease), given the MOI gene model assumed.

Table 7.1. Relative risk assumptions, by mode of inheritance

Inheritance model	Major homozygote risk Ψ_{AA}	Minor homozygote risk $\Psi_{aa} = \frac{\Pr(\text{case}/aa)}{\Pr(\text{case}/AA)}$	Heterozygote risk $\Psi_{Aa} = \frac{\Pr(\text{case}/aA)}{\Pr(\text{case}/AA)}$
Recessive	1	Φ	1
Dominant	1	Φ	Φ
Additive	1	$2 \times \Phi - 1$	Φ
Multiplicative	1	$\Phi \times \Phi$	Φ

Pr = probability. Φ = genetic inheritance risk.

Source: Iles (2002).[22]

This genotype-specific process can be represented by the following logic:

- Major homozygote (**AA**)

  If the **AA** (nondisease) genotype is selected, the probability of a case equals the disease penetrance, P_j.

- Minor homozygote (**aa**)

  $\boldsymbol{\Psi}_{\mathbf{aa}}$ is the exogenous risk and represents the ratio of two probabilities: the probability of a case for a minor homozygote divided by the probability (Pr) of a case for a major homozygote. In other words,

  $$\boldsymbol{\Psi}_{\mathbf{aa}} = \Pr(\text{case}/\mathbf{aa}) \,/\, \Pr(\text{case}/\mathbf{AA}) = x/P_j. \tag{7.1}$$

  Thus, the probability of a case given the minor genotype is

  $$x = \boldsymbol{\Psi}_{\mathbf{aa}} \times P_j. \tag{7.2}$$

- Heterozygote (**aA**)

  By the same argument, the phenotype risk given a heterozygote is

  $$\boldsymbol{\Psi}_{\mathbf{aA}} = \Pr(\text{case}/\mathbf{aA}) \,/\, \Pr(\text{case}/\mathbf{AA}) = y/P_j. \tag{7.3}$$

  Thus, the risk of a case given the heterozygote genotype is

  $$y = \boldsymbol{\Psi}_{\mathbf{aA}} \times P_j, \tag{7.4}$$

  where $\boldsymbol{\Psi}_{\mathbf{aA}}$ is the assumed risk factor and P_j is the assumed penetrance.

Implicit in equations (7.1) through (7.4) is a consistent definition of penetrance defined as the proportion of cases that are present in the major genotype **AA**.

Using the estimate of x from equation (7.2) and y from equation (7.4), we specified a subject as a case (1) or control (0) at random using the four different MOI models from Table 7.1. For the MOI models that assume an elevated risk from the minor and the heterozygote genotypes, we would expect a higher proportion of cases to be more easily identified via the statistical procedures. Specifying risk depends on known and unknown disease mechanisms. Some consider a relative risk of 1.7 high and a risk of 1.3 to be a more realistic assumption for complex diseases.[25,26] We limited our focus to relative risks in the range of 1.1 to 1.2.

Note that we assigned cases and controls so that there would be no possibility for the introduction of bias. We chose to ignore errors in both genotype and the phenotype measurements, which in a real experiment could be a source of bias (we examined both sources of error in an earlier study[20]). This process continued until we created *n1* cases and *n2* controls. We then applied a set of statistical methods (identified subsequently) to predict associations, and recorded and tracked the results. For each set of unique factor combinations (i.e., penetrance, sample sizes, relative risk levels, and MOI categories), we generated 1,000 replicate experiments.

Exogenously, we specified the genetic inheritance (GI) relative risk of disease as 1.10, 1.15, and 1.20 and defined it in the overview as the ratio of the probability of a disease diagnosis for subjects, dividing the wild-type gene by the probability of disease, based on all genetic and nongenetic causes. We also defined a second relative risk component based on a specific environmental exposure (EE). We defined this ratio as the probability of a disease given the EE divided by the probability of a diagnosis given no EE. In discussion of these experiments, we use the notation Φ to represent GI and Π to represent EE.

The form of the EE relative risk can be specified using a variety of assumptions. In all scenarios, the genetic risk is first used to determine the phenotype status (case or control). Then the environmental risk calculation determines whether the phenotype status is altered from control to case according to the EE assumptions. We assume that the form of the EE effect is not known but that the specific variable is known. In the following experiments, we use E = age as a proxy for the different assumed forms of exposure, and we assign E a value obtained from a uniform distribution of

30 to 70. The value of E controls the EE risk according to different experiment designs. The main objective of this assessment is to identify whether one statistical model outperforms all other models and how much variation occurs across the different experiments.

For all experiments, we used the GI as described previously.

Experiment 1—The Main Effects Model. For the first experiment, half of the population (selected at random and assigned $50 < E < 71$) incurred an EE relative risk (Π). The assigned risk value was 1.10, 1.20, 1.30, or 1.40. The other half of the population (assigned $29 < E < 51$) incurred no risk ($\Pi = 1.0$). Thus, Experiment 1 simulates a fixed EE. When the determinant risk variable, E, exceeds a threshold, a positive diagnosis is more likely to occur. This is identified as the *fixed risk, main effects, no interaction model.*

Experiment 2—The Interaction Effects Model. For the second scenario, again half of the population (selected at random and assigned $50 < E < 71$) incurred an EE relative risk (Π). This risk value was 1.10, 1.20, 1.30, or 1.40, but only if the subject also had a wild-type allele (i.e., a heterozygote or minor homozygote genotype). The other component of the population ($50 < E < 71$ and genotype = **AA**) incurred no EE risk ($\Pi = 1.0$). Experiment 2 also simulates a fixed EE but only if the genotype contains a wild-type allele. This is identified as the *fixed-risk, main effect with interaction model.*

Experiment 3—The Main Effects Log-Linear Risk Model. For the third scenario, the entire population (randomly assigned $30 \leq E \leq 70$) incurred an EE relative risk (Π) which was related to E in the following manner:

$y = (E - 30)/40$.

$\Pi = X^y$ (X to the y power), where $X = \{1.10, 1.20, 1.30, 1.40\}$.

Experiment 3 simulates a *log-linear variable risk model*, with larger values of E conveying additional risk levels. As in Experiment 1, there is no interaction between the GI and EE risks.

Experiment 4—The Interaction Effects Log-Linear Risk Model. The fourth scenario is the same as the third scenario, but the risk applies only if the subject has a wild-type allele.

Experiment 4 simulates a variable-risk scenario with larger values of E conveying higher risk levels—but only if the genotype contains a wild-type allele. This is the *log-linear variable risk main effect with genotype interaction model.*

For each experiment type, we varied the gene model to determine the relative power differences across model specification. Overall, Experiment 1 data have a step function relationship to EE and no interaction or difference in slopes (or EE step heights) across the three genotypes. In contrast, the Experiment 2 data has a step function relationship with EE where the **aa** and **aA** genotypes have the same slope (step height) but different intercepts. The **AA** genotype relationship to the EE is flat or has zero slope (no step up). In Experiment 3, the relationship to EE is log-linear, with equal slopes for all three genotypes. Finally, in Experiment 4, the relationship to EE is log-linear; the **aa** and **aA** genotypes have the same slope but different intercepts; and the **AA** genotype relationship to the EE is flat, or has zero slope.

Statistical Models

All models tested assumed a logistic regression (LR) specification. This form is commonly used in association studies involving environmental interactions.[21]

Table 7.2 shows the variables used in the different models.

Table 7.2. Descriptions of variables used in the logistic regression models

Variable category	Name	Form	Values
Genotype	G	Continuous	0, 1, 2
Genotype	g1, g2	Categorical	0, 1
Environmental	E	Continuous	30–70
Environmental	e1	Categorical	0, 1
Interaction	g1 × E	Mixed	0, 30–70
Interaction	g1 × e1, g2 × e1	Categorical	0, 1

Notes: G = the number of wild-type alleles for the genotype (0, 1, 2).

g1 = 1 if the subject's genotype is a heterozygote, otherwise g1 = 0.

g2 = 1 if the subject's genotype is a minor or wild homozygote, otherwise g2 = 0.

E = a variate from a uniform distribution (30–70) that suggests it is an age.

e1 = an indicator variable set to 0 if E < 50. Otherwise e1 = 1.

The difference between the experiments is straightforward. For subjects younger than 50 years old, there is no risk from environmental exposure (i.e., the relative risk = 1.0) in Experiments 1 and 2; subjects older than 50 have an environmental exposure (i.e., the risk is greater than 1.0). However, in Experiment 2, only subjects aged 50 or older who have a wild-type allele are assumed to have the assigned risk. The main discriminator between Experiments 1 and 3 (and Experiments 2 and 4) is the risk characterization. For

Experiments 1 and 2, the risk is intended to be an all-or-nothing process akin to a toxic exposure that occurs sometime after the subject reaches age 50. For Experiments 3 and 4, the risk due to an environmental exposure is present in all subjects and increases as age increases. Table 7.3 summarizes the experiments.

Table 7.3. Experiment description

Experiment	Risk type	Scenario description	Exposure action
1	Fixed EE	Chemical exposure	Risk applies to half of the population
2	Fixed EE with interaction	Chemical exposure affects genotype	Risk applies to the half of the population who have the wild-type allele
3	Variable EE	Advancing age	Risk applies to half of the population and increases with age
4	Variable EE with interaction	Advancing age affects genotype	Risk applies to the half of the population who have the wild-type allele, and risk increases with age

EE = environmental exposure.

We used three specific statistical models to assess the data generated by the four experiments. Each assumed an intercept term and had the following form:

- Model 1 is a logistic regression model with a single variable genotype (G) main effect (2*df*). This is a candidate model if no environmental exposure were suspected.
- Model 2 is a logistic regression mixed main effects and interaction model (g1, g2, E, g1 × E, g2 × E) (6*df*). This is a fully specified model that assumes that the environmental exposure is a continuous variable.
- Model 3 is also a logistic regression mixed main effects and interaction model (g1, g2, e1, g1 × e1, g2 × e1) (6*df*). This is the fully specified categorical model and assumes that the environmental exposure has a specific (all-or-nothing) categorical variable form.

Table 7.4 summarizes the specific regression models we used in this study. Note that we initially compared six models. Two were gene-only models—a 1*df* (log-additive test) and a 2*df* test—and four were main effects plus interaction models. We had two environmental exposure specifications (E and e1) and two genetic inheritance specifications (G and g1, g2). From the six initial models,

we selected the three models that dominated the other three: M-1, M-2, and M-3. We dropped the other three models (M-1a, M-2a, and M-3a) from our assessment.

Table 7.4. Statistical models assessed

Model	Main effects	Interactions	*df*	Test statistic
M-1	G	NA	1	LLH[log(α, G)] – LLH[log(α)]
M-1a	g1, g2	NA	2	LLH[log(α, g1, g2)] – LLH[log(α)]
M-2	g1, g2, E	g1 × E, g2 × E	5	LLH[log(α, g1, g2, E, g1*E, g2*E)] – LLH[log(α)]
M-2a	G, E	G × E	3	LLH[log(α, G, E, G × E)] – LLH[log(α)]
M-3	g1, g2, e1	g1 × e1, g2 × e1	5	LLH[log(α, g1, g2, e1, g1*e1, g2*e1)] – LLH[log(α)]
M-3a	G, e1	G × e1	3	LLH[log(α, G, e1, G × e1)] – LLH[log(α)]

df = degrees of freedom; NA = not applicable; LLH = log-likelihood. α = the logit scale intercept for the line relating environmental exposure (EE) to the log-odds risk among those subjects with the nondisease genotype **AA**.

Notes: G = the number of wild-type alleles for the genotype(0, 1, 2).

g1 = 1 if the subject's genotype is a heterozygote, otherwise g1 = 0.

g2 = 1 if the subject's genotype is a minor or wild homozygote, otherwise g2 = 0.

E = age, 30–70.

The test statistics we used in our analyses are defined as the difference between two log-likelihood (LLH) statistics. The first is specific to the model used, and the second is based on a model with only the intercept term.

Results

Association Analysis

In this section, we describe the power profiles that result by applying the models described in Table 7.4 to the data generated according to the four different experiments in Table 7.3. We focus on detecting the associations between the combined genotype-environmental factors on phenotype outcome (disease diagnosis). We assess the importance of model specification in predicting the presence of association with a phenotype of interest and to what degree the gene model and genotype environment interactions influence power. In the following section on genotype associations, we assess the role

of the genotype alone in predicting association while controlling for the environmental influence.

Note that in calculating all power results in this section, we assumed that the Type I error rate was 10^{-8}. However, because all combined environmental exposure and genetic inheritance risk values are greater than 1.0 in all of our experiments, only Type II errors were possible.

Table 7.5 shows the data generated using the protocol for Experiment 1. Note that for this and all subsequent tables in this section, the highest power value for each risk profile within the three MOI categories is bolded to highlight the optimal model. For each genetic inheritance (GI) risk level (Φ) there is an environmental exposure (EE) risk level (Π) equal to 1.0, indicating no EE risk.

Table 7.5. Power values, by statistical model, Φ, and Π: Experiment 1—all gene models

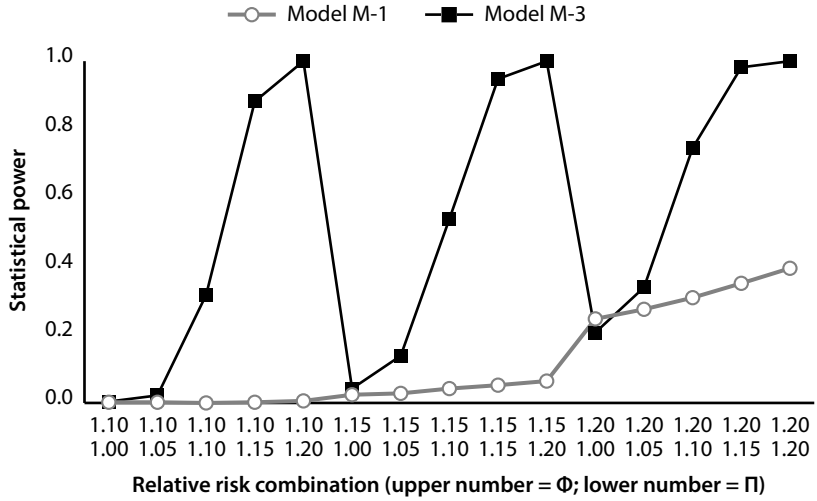
		Additive			Dominant			Recessive		
Φ	**Π**	**M-1**	**M-2**	**M-3**	**M-1**	**M-2**	**M-3**	**M-1**	**M-2**	**M-3**
1.10	1.00	.002	.004	**.004**	.000	.004	**.004**	.000	.000	**.000**
1.10	1.05	.002	.016	**.022**	.000	.040	**.044**	.000	.000	**.000**
1.10	1.10	.000	.160	**.316**	.000	.214	**.354**	.000	.050	**.122**
1.10	1.15	.002	.654	**.882**	.000	.728	**.924**	.000	.488	**.810**
1.10	1.20	.006	.986	**1.00**	.000	.992	**1.00**	.000	.968	**1.00**
1.15	1.00	.024	.046	**.042**	.000	.028	**.024**	.000	.000	**.000**
1.15	1.05	.028	.102	**.138**	.000	.082	**.102**	.000	.000	**.000**
1.15	1.10	.042	.378	**.538**	.000	.336	**.468**	.000	.038	**.144**
1.15	1.15	.052	.838	**.948**	.000	.832	**.954**	.000	.528	**.806**
1.15	1.20	.064	.996	**1.00**	.000	.994	**1.00**	.000	.958	**.998**
1.20	1.00	.246	.210	**.206**	.004	.100	**.104**	.000	.000	**.000**
1.20	1.05	.274	.308	**.338**	.002	.214	**.240**	.000	.002	**.004**
1.20	1.10	.308	.630	**.746**	.006	.552	**.684**	.000	.054	**.142**
1.20	1.15	.350	.908	**.982**	.016	.912	**.972**	.002	.586	**.852**
1.20	1.20	.394	.996	**1.00**	.028	.994	**1.00**	.004	.972	**.998**

Φ = genetic inheritance (GI) risk level; Π = environmental exposure (EE) risk level.
Note: Table 7.4 defines the statistical models (M-1, M-2, and M-3).

Figure 7.1 shows the data generated using the protocol for Experiment 1 for the additive gene model. Figure 7.1 includes the optimal model (M-3, identified by the boldfaced cells in Table 7.5) and the model that does not include an EE variable in its specification (M-1). The results presented in

Table 7.5 and Figure 7.1 indicate that there is little difference in performance between models when the risk of EE is not present.

Figure 7.1. Power curves, by statistical model, Φ, and Π: Experiment 1—additive gene model



Φ = genetic inheritance (GI) risk level; Π = environmental exposure (EE) risk level.
Note: Table 7.4 defines the statistical models (M-1 and M-3).

The results that Figure 7.1 and Table 7.5 show indicate the following:

- The power profile of M-1 is substantially less than that of models M-2 and M-3. M-1 represents a typical single-locus method used in a GWAS that ignores environmental influences. We conclude that not including an EE reduces the likelihood of the locus being associated with the phenotype.
- M-3 is the most powerful of the three models. This is expected because the Experiment 1 protocol should generate data consistent with the M-3 model formulation.
- The difference between the profiles of M-2 and M-3 is a result of the manner used to characterize the EE functional form. Because the data were generated in a manner compatible with the e1 variable used in M-3, that model generated more accurate power predictions.

Note that in the full M-3, the overall intercept is the log of the intercept for the line that relates EE to the log-odds risk among those subjects with the

nondisease genotype **AA**. The coefficient associated with the g1 main effect is testing for the difference between intercepts for the subjects with genotype **aA** and those with genotype **AA**. Similarly, the g2 main effect coefficient is testing for the difference between the intercepts for subjects with the **aa** genotype and those with the **AA** genotype.

The EE main effect coefficient is the height of the step in the step function relating EE to log-odds-risk for subjects with the **AA** genotype, and it, therefore, tests for a common EE step height across all three genotypes. The g1 × E interaction coefficient is the difference between the step heights for the **aA** subjects and the **AA** subjects. Similarly, the g2 × E coefficient is the difference between the step heights for the **aa** subjects and the **AA** subjects. Because the **AA**, **aA**, and **aa** step heights/slopes associated with the EE environmental effect are all equal in Experiments 1 and 3, only the common main effect (ME) associated with EE contributes to association prediction in those data sets, and the interaction terms are superfluous.

Table 7.6 and Figure 7.2 show the results of applying the three models that Table 7.4 describes to the data generated according to the Experiment 2 protocol (see Table 7.3). The results from Experiment 2 indicate the following:

- Although M-1 does not adjust for EE, the observed (relatively) high power profiles for high EE risk levels suggest that the GI-EE interaction effect is embedded in the M-1 power values, and the high power profiles are credited as a genotype main effect.
- As in Experiment 1, M-3 outperforms all other models because the variable e1 properly characterizes EE behavior. This clearly demonstrates the value of preprocessing (i.e., mining) the data before committing to a specific association model.

Table 7.7 and Figure 7.3 show the results of applying the three statistical models described in Table 7.4 to the data generated according to the Experiment 3 protocol (see Table 7.3). The results from Experiment 3 indicate the following:

- M-1 consistently performs below M-2 and M-3, indicating that not including an EE term limits the association assessment.
- In general, model M-2 produces better power profiles than M-3. This is expected given that the EE incremental risk is linearly related to the log of EE. Thus, model M-2 is more consistent with the protocol used to generate the data in Experiment 3.

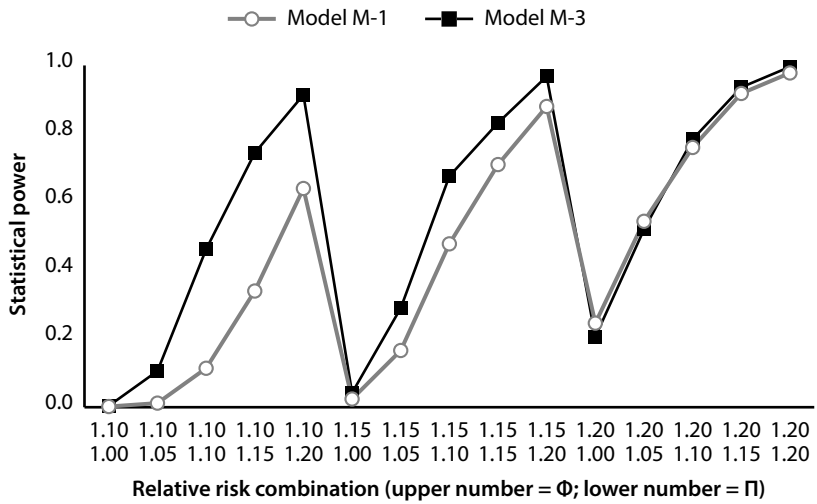
Table 7.6. Power values, by statistical model, Φ, and Π: Experiment 2—all gene models

		Additive			Dominant			Recessive		
Φ	**Π**	**M-1**	**M-2**	**M-3**	**M-1**	**M-2**	**M-3**	**M-1**	**M-2**	**M-3**
1.10	1.00	.002	.004	**.004**	.000	.004	**.004**	.000	.000	**.000**
1.10	1.05	.012	.086	**.106**	.000	.086	**.094**	.000	.000	**.002**
1.10	1.10	.114	.394	**.462**	.010	.408	**.442**	.002	.072	**.116**
1.10	1.15	.340	.702	**.744**	.078	.690	**.738**	.022	.376	**.462**
1.10	1.20	.640	.856	**.914**	.318	.844	**.896**	.104	.642	**.714**
1.15	1.00	.024	.046	**.042**	.000	.028	**.024**	.000	.000	**.000**
1.15	1.05	.166	.258	**.290**	.006	.212	**.222**	.000	.000	**.004**
1.15	1.10	.478	.634	**.676**	.082	.570	**.590**	.002	.090	**.126**
1.15	1.15	.710	.808	**.832**	.380	.846	**.872**	.082	.398	**.464**
1.15	1.20	.880	.952	**.968**	.730	.934	**.954**	.230	.654	**.740**
1.20	1.00	.246	.210	**.206**	.004	.100	**.104**	.000	.000	**.000**
1.20	1.05	.544	.518	**.520**	.084	.404	**.426**	.008	.002	**.006**
1.20	1.10	.760	.776	**.784**	.336	.752	**.784**	.052	.120	**.162**
1.20	1.15	.918	.926	**.936**	.736	.920	**.944**	.184	.418	**.496**
1.20	1.20	.978	.990	**.996**	.916	.984	**.986**	.345	.640	**.725**

Φ = genetic inheritance (GI) risk level; Π = environmental exposure (EE) risk level.

Note: Table 7.4 defines the statistical models (M-1, M-2, and M-3).

Figure 7.2. Power curves, by statistical model, Φ, and Π: Experiment 2—additive gene model



Φ = genetic inheritance (GI) risk level; Π = environmental exposure (EE) risk level.

Note: Table 7.4 defines the statistical models (M-1 and M-3).

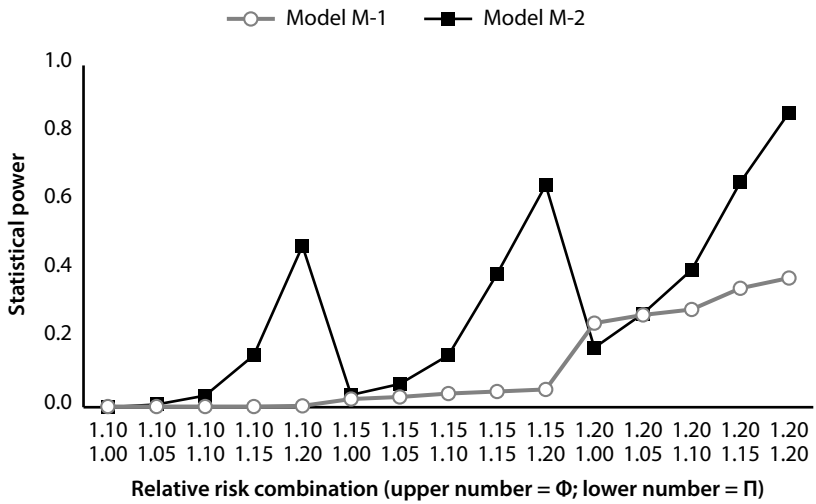
Table 7.7. Power values, by statistical model, Φ, and Π: Experiment 3—all gene models

		Additive			Dominant			Recessive		
Φ	**Π**	**M-1**	**M-2**	**M-3**	**M-1**	**M-2**	**M-3**	**M-1**	**M-2**	**M-3**
1.10	1.00	.002	.000	.004	.000	**.006**	.004	.000	.000	.000
1.10	1.05	.002	**.008**	.006	.000	**.014**	.014	.000	.000	.000
1.10	1.10	.002	**.034**	.030	.000	**.044**	.030	.004	.004	.000
1.10	1.15	.002	**.152**	.096	.000	**.190**	.120	.062	.024	.018
1.10	1.20	.004	**.472**	.296	.000	**.500**	.260	.258	.228	.066
1.15	1.00	.024	.036	**.042**	.000	**.034**	.024	.000	.000	.000
1.15	1.05	.030	**.068**	.058	.000	.048	**.052**	.000	.000	.000
1.15	1.10	.040	**.152**	.106	.000	**.162**	.118	.002	.002	.002
1.15	1.15	.046	**.390**	.278	.000	**.384**	.302	.048	.034	.012
1.15	1.20	.052	**.650**	.524	.000	**.670**	.508	.308	.278	.074
1.20	1.00	.246	.174	**.206**	.004	**.114**	.104	.000	.000	.000
1.20	1.05	.270	**.274**	.250	.002	**.166**	.150	.000	.000	.000
1.20	1.10	.286	**.402**	.376	.002	**.344**	.260	.004	.002	.002
1.20	1.15	.348	**.658**	.548	.012	**.520**	.490	.098	.058	.032
1.20	1.20	.378	**.862**	.718	.024	**.796**	.660	.299	.289	.107

Φ = genetic inheritance (GI) risk level; Π = environmental exposure (EE) risk level.

Note: Table 7.4 defines the statistical models (M-1, M-2, and M-3).

Figure 7.3. Power curves, by statistical model, Φ, and Π: Experiment 3—additive gene model



Φ = genetic inheritance (GI) risk level; Π = environmental exposure (EE) risk level.

Note: Table 7.4 defines the statistical models (M-1 and M-2).

Table 7.8 and Figure 7.4 show the results for Experiment 4. They indicate the results of applying the three models described in Table 7.4 to the data generated according to the Experiment 4 protocol (see Table 7.3).

Table 7.8. Power values, by statistical model, Φ, and Π: Experiment 4—all gene models

		Additive			Dominant			Recessive		
Φ	**Π**	**M-1**	**M-2**	**M-3**	**M-1**	**M-2**	**M-3**	**M-1**	**M-2**	**M-3**
1.10	1.00	.002	.006	.004	.000	.006	.004	.000	.000	.000
1.10	1.05	.008	.054	.006	.000	.068	.054	.000	.000	.002
1.10	1.10	.078	.236	.030	.006	.284	.226	.002	.014	.008
1.10	1.15	.220	.478	.096	.048	.500	.476	.008	.110	.078
1.10	1.20	.510	.672	.678	.148	.624	.632	.046	.314	.294
1.15	1.00	.024	.074	.042	.000	.046	.024	.000	.000	.000
1.15	1.05	.118	.208	.164	.002	.162	.130	.000	.000	.000
1.15	1.10	.384	.514	.476	.040	.454	.410	.000	.018	.016
1.15	1.15	.618	.688	.658	.232	.698	.684	.046	.162	.132
1.15	1.20	.820	.830	.838	.570	.802	.792	.144	.354	.328
1.20	1.00	.246	.250	.206	.004	.138	.104	.000	.000	.000
1.20	1.05	.464	.466	.406	.046	.354	.290	.004	.002	.000
1.20	1.10	.714	.714	.692	.222	.642	.624	.026	.040	.022
1.20	1.15	.892	.876	.864	.622	.816	.824	.136	.216	.170
1.20	1.20	.954	.940	.944	.848	.916	.930	.257	.317	.343

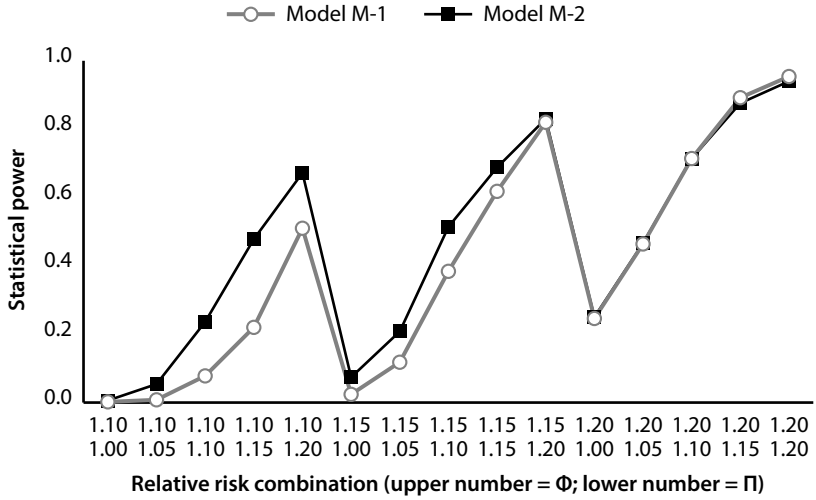
Φ = genetic inheritance (GI) risk level; Π = environmental exposure (EE) risk level.

Note: Table 7.4 defines the statistical models (M-1, M-2, and M-3).

Table 7.8 and Figure 7.4 show results that indicate the following:

- Consistent with Experiment 2's results, model M-1 does not adjust for EE, but because of the influence of GI-EE interaction effects, M-1 displays higher power profiles for large EE risk levels.
- As in Experiment 3, M-2 outperforms M-3 because it better characterizes the EE by using the variable E (age) and further demonstrates the value of preprocessing (i.e., mining) the data before committing to a specific association model.
- In the presence of GI-EE interaction effects, the genetic-only model (M-1) performs better than anticipated.

Figure 7.4. Power curves, by statistical model, Φ, and Π: Experiment 4—additive gene model



Φ = genetic inheritance (GI) risk level; Π = environmental exposure (EE) risk level.
Note: Table 7.4 defines the statistical models (M-1 and M-2).

Genotype Associations

The analysis in the previous section focused exclusively on composite associations—that is, whether a specific gene plus an environmental factor associates with a phenotype. As we noted earlier, our main interest was separating main genetic effects from environmental effects and their interactions. To accomplish this, we defined a total effect test (TOT) that adjusts for EE where

$$\text{TOT} = \text{LLH}\,[\log(\alpha, g1, g2, e1, g1 \times e1, g2 \times e1)] - \text{LLH}\,[\log(\alpha, e1)] \quad (7.5)$$

is the test we applied to the data generated by Experiment 1 and 2 protocols, and

$$\text{TOT} = \text{LLH}\,[\log(\alpha, g1, g2, E, g1 \times E, g2 \times E)] - \text{LLH}\,[\log(\alpha, E)] \quad (7.5a)$$

is the test that we applied to the data generated by Experiment 3 and 4 protocols.

TOT is the association test that measures genetic effects (main and interactive) and is adjusted for the environmental effect.[27] TOT simultaneously measures whether the **aA** and **aa** intercepts are different from the **AA** intercept and whether the **aA** and **aa** slopes are nonzero, given that the **AA** slope on EE is zero. This test was used to test for association from all causes.

We also define two additional tests for genotype-environment interactions, INT, as follows:

$$\text{INT} = \text{LLH}\,[\log(\alpha, e1, g1, g2, g1 \times e1, g2 \times e1)] - \text{LLH}\,[\log(\alpha, e1, g1, g2)] \quad (7.6)$$

and

$$\text{INT} = \text{LLH}\,[\log(\alpha, E, g1, g2, g1 \times E, g2 \times E)] - \text{LLH}\,[\log(\alpha, E, g1, g2)]. \quad (7.6a)$$

The INT test subtracts the main effects for g1, g2, and EE from the TOT and tests whether the EE steps (or slopes) for the **aA** and **aa** genotypes are different from the corresponding EE step (slope) for genotype **AA**.

The final test measures the influence of the genetic main effects (ME).

$$\text{ME} = \text{LLH}\,[\log(\alpha, e1, g1, g2)] - \text{LLH}\,[\log(\alpha, e1)] \quad (7.7)$$

is the test applied to the data generated by Experiments 1 and 2 protocols and

$$\text{ME} = \text{LLH}\,[\log(\alpha, E, g1, g2)] - \text{LLH}\,[\log(\alpha, E)] \quad (7.7a)$$

is the corresponding test for data from Experiments 3 and 4 protocols.

The ME tests check whether the estimated **aA** and **aa** intercepts differ from the **AA** intercept, conditioned on the EE step sizes (e1 in experiments 1and 2) or the EE slopes (E in experiments 3 and 4) being equal for all three genotypes.

Note that for Experiments 2 and 4, both the **AA** step (coefficient of e1) and slope (coefficient of E) on EE are zero, and therefore the coefficient for the EE main effect (assuming that the M-3 is operating) is estimating zero; the two interaction columns are estimating the **aA** step/slope minus zero and the **aa** step/slope minus zero, respectively.

Typically, these three tests would be applied sequentially: TOT followed by INT, then ME. Assessing whether an interactive or noninteractive genetic association is obtained would depend on the result of the preceding test.

For example, if TOT is nonsignificant, the process stops, and we conclude that there is no connection between the genetic locus and the phenotype. Otherwise, we would apply the INT test. If INT was significant, we could conclude that the locus and the phenotype are significantly related, with the caveat that the strength of the genotype effect varies by the EE risk level. The ME test would only be applied if the TOT is significant and INT test is not significant. In this case, the ME test would be applied to affirm that the genetic and environmental effects are operating independently of each other and to assert that a common genotype main effect exists that applies to all EE levels. Tables 7.9 through 7.13 show the results of running the three tests (TOT, INT, and ME).

Table 7.9. Total effects test (TOT) power values, by risk profile, Φ, and Π—all experiments and gene models, *N* = 200,000

		Experiment 1			Experiment 2			Experiment 3			Experiment 4		
Φ	Π	TOT^ Rec	TOT^ Dom	TOT^ Add	TOT^ Rec	TOT^ Dom	TOT^ Add	TOT* Rec	TOT* Dom	TOT* Add	TOT* Rec	TOT* Dom	TOT* Add
1.10	1.00	.319	.601	.574	.319	.601	.574	.328	.573	.538	.328	.573	.538
1.10	1.05	.328	.602	.612	.523	.777	**.827**	.340	.619	.608	.719	**.958**	**.968**
1.10	1.10	.343	.604	.577	**.806**	**.949**	**.953**	.377	.636	.596	**.990**	**1.00**	**1.00**
1.10	1.15	.344	.613	.625	**.958**	**.994**	**.999**	.408	.684	.662	**1.00**	**1.00**	**1.00**
1.10	1.20	.358	.622	.637	**.993**	**.999**	**1.00**	.492	.709	.704	**1.00**	**1.00**	**1.00**
1.15	1.00	.341	.725	.739	.341	.725	.739	.336	.725	.713	.336	.725	.713
1.15	1.05	.363	.745	.769	.534	**.918**	**.919**	.367	.750	.766	.764	**.995**	**.997**
1.15	1.10	.375	.770	.773	**.837**	**.986**	**.993**	.427	.796	**.828**	**.992**	**1.00**	**1.00**
1.15	1.15	.358	.751	.776	**.934**	**.999**	**1.00**	.480	**.832**	**.870**	**1.00**	**1.00**	**1.00**
1.15	1.20	.351	.775	.787	**.995**	**.999**	**1.00**	.487	**.861**	**.893**	**1.00**	**1.00**	**1.00**
1.20	1.00	.444	**.889**	**.915**	.444	**.889**	**.915**	.439	**.851**	**.904**	**.439**	**.851**	**.904**
1.20	1.05	.456	**.900**	**.918**	.631	**.986**	**.982**	.490	**.912**	**.950**	**.809**	**1.00**	**.999**
1.20	1.10	.477	**.897**	**.943**	**.854**	**.998**	**1.00**	.514	**.949**	**.949**	**.995**	**1.00**	**1.00**
1.20	1.15	.471	**.917**	**.935**	**.973**	**1.00**	**1.00**	.594	**.946**	**.975**	**1.00**	**1.00**	**1.00**
1.20	1.20	.514	**.925**	**.945**	**.996**	**1.00**	**1.00**	.632	**.961**	**.986**	**1.00**	**1.00**	**1.00**

Φ = genetic inheritance (GI) risk level; Π = environmental exposure (EE) risk level; TOT = total effects test; gene models: Rec = recessive, Dom = dominant, Add = additive.

Note: ^ = TOT from equation 5; * = TOT from equation 5a; $\alpha \leq 10^{-8}$.

Consider that every replicate in every cell produced by the simulation experiments is designed to generate a genotype-phenotype association (albeit at low risk). Some of these replicates influenced by an EE also contribute toward association. However, in a perfect statistical world, all are generated to predict an association with the phenotype. The fact that they do not indicates the limitations of the GWAS process.

In addition, Table 7.9 suggests the following:

- Association detection involving recessive genes is difficult to identify and accordingly requires a larger sample size than we used in our experiments.
- Scenarios involving gene-environment interactions (Experiments 2 and 4) greatly influence whether genetic influences can be detected by a gene-only model.

- The type of EE process influences the ability to detect an association, whether the effect is caused by a chemical-type exposure (Experiments 1 and 2) or by aging (Experiments 3 and 4).

Table 7.10 presents the results of applying the INT test to all experiments and all gene models. Not shown are the results for Experiments 1 and 3, which generated data without interaction effects. They estimate no interaction between GI and EE (as they should), so those results are not shown. Note that the Type 1 α thresholds in Table 7.11 for generating power estimates for all cells are $\leq 10^{-2}$.

Table 7.10. Genotype-environment interactions (INT) power values, by risk profile (Φ and Π)—all experiments and gene models, *N* = 10,000

		Experiment 2			Experiment 4		
Φ	Π	INT^ Rec	INT^ Dom	INT^ Add	INT* Rec	INT* Dom	INT* Add
1.10	1.00	.319	.601	.574	.319	.601	.574
1.10	1.05	.328	.602	.612	.523	.777	.827
1.10	1.10	.343	.604	.577	.806	.949	.953
1.10	1.15	.344	.613	.625	.958	.994	.999
1.10	1.20	**.358**	**.622**	**.637**	.993	.999	1.00
1.15	1.00	.341	.725	.739	.341	.725	.739
1.15	1.05	.363	.745	.769	.534	.918	.919
1.15	1.10	.375	.770	.773	.837	.986	.993
1.15	1.15	.358	.751	.776	.934	.999	1.00
1.15	1.20	**.351**	**.775**	**.787**	.995	.999	1.00
1.20	1.00	.444	.889	.915	.444	.889	.915
1.20	1.05	.456	.900	.918	.631	.986	.982
1.20	1.10	.477	.897	.943	.854	.998	1.00
1.20	1.15	.471	.917	.935	.973	1.00	1.00
1.20	1.20	**.514**	**.925**	**.945**	.996	1.00	1.00

Φ = genetic inheritance (GI) risk level; Π = environmental exposure (EE) risk level; INT = genotype-environment interactions; gene models: Rec = recessive, Dom = dominant, Add = additive.

Note: ^ = INT from equation 6; * = INT from equation 6a; $\alpha \leq 10^{-2}$.

Table 7.10 reaffirms the results of Table 7.9:

- Power values for recessive genes are very low and accordingly were more difficult to identify than other gene models.

- Gene-environment interactions influence association outcomes. This is evidenced by all cells of the no-interaction experiments (Experiments 1 and 3) having power values <.004.
- The type of EE process influences the detection of an association, whether the effect is due to an exposure (Experiments 1 and 2) or is due to an aging mechanism (Experiments 3 and 4).
- Interaction effects achieve significant levels in Experiment 2 for risk values of EE ≥ 1.2 only.

Table 7.11. Main effects (ME) power values, by risk profile (Φ and Π)—all gene models, *N* = 10,000

		Experiment 1			Experiment 3		
Φ	Π	ME^ Rec	ME^ Dom	ME^ Add	ME* Rec	ME* Dom	ME* Add
1.10	1.00	.123	.463	.434	.139	.450	.417
1.10	1.05	.131	.478	.476	.156	.479	.437
1.10	1.10	.138	.491	.459	.138	.476	.399
1.10	1.15	.138	.523	.500	.149	.498	.478
1.10	1.20	.154	.507	.495	.167	.506	.487
1.15	1.00	.153	.626	.628	.147	.608	.585
1.15	1.05	.179	.642	.664	.146	.619	.659
1.15	1.10	.193	.676	.660	.177	.654	.675
1.15	1.15	.163	.670	.670	.191	.653	.675
1.15	1.20	.171	.668	.682	.182	.676	.695
1.20	1.00	.204	**.805**	**.850**	.239	.755	**.834**
1.20	1.05	.244	**.824**	**.850**	.236	**.807**	**.861**
1.20	1.10	.250	**.829**	**.887**	.239	**.817**	**.866**
1.20	1.15	.264	**.840**	**.873**	.274	**.816**	**.867**
1.20	1.20	.315	**.854**	**.893**	.268	**.830**	**.878**

Φ = genetic inheritance (GI) risk level; Π = environmental exposure (EE) risk level; ME = main effects; gene models: Rec = recessive, Dom = dominant, Add = additive.

Note:^ = ME from equation 7; * = ME from equation 7a; $\alpha \leq 10^{-2}$.

Table 7.11 presents the results of the ME test for Experiments 1 and 3. ME results for Experiments 2 and 4 are not shown because they were generated by a protocol that produced EE and GI interactions, and if the INT test demonstrated significance (as it should have), the ME tests would have been unnecessary. In all cases, the alpha threshold was set to 10^{-2}.

Table 7.11 reaffirms the results of Tables 7.9 and 7.10:

- Associations involving recessive genes are more difficult to identify.
- Gene-environment interactions influence association outcomes.
- The type of process influences the detection of an association, as shown by differences between power values for exposure mechanisms such as those resembling chemical spills (Experiments 1 and 2) and those recognizing aging mechanisms (Experiments 3 and 4).
- Main effects are only considered significant for larger risk values of genetic inheritance (those with a risk of 1.2 or greater).

Note that the power threshold values are set to low ($\alpha \leq 10^{-2}$) for the interaction and ME tables. To investigate the effect of a very large N, we repeated the simulation process with N =200,000 and reduced the threshold to ($\alpha \leq 10^{-8}$). Tables 7.12 and 7.13 present the results.

Table 7.12. Genotype-environment interactions (INT) power values, by risk profile (Φ and Π)—all gene models, *N* = 200,000

		Experiment 2			Experiment 4		
Φ	Π	INT^ Rec	INT^ Dom	INT* Add	INT* Rec	INT* Dom	INT* Add
1.10	1.00	.00	.00	.00	.00	.00	.00
1.10	1.05	.00	.00	.03	.00	.00	.00
1.10	1.10	.75	.77	.88	.35	.50	.56
1.10	1.15	1.0	1.0	1.0	1.0	1.0	1.0
1.10	1.20	1.0	1.0	1.0	1.0	1.0	1.0
1.15	1.00	.00	.00	.00	.00	.00	.00
1.15	1.05	.00	.01	.03	.00	.00	.01
1.15	1.10	.78	.86	.91	.44	.52	.60
1.15	1.15	1.0	1.0	1.0	.96	1.0	1.0
1.15	1.20	1.0	1.0	1.0	1.0	1.0	1.0
1.20	1.00	.00	.00	.00	.00	.00	.00
1.20	1.05	.01	.04	.05	.00	.00	.00
1.20	1.10	.75	.92	.96	.45	.65	.74
1.20	1.15	.99	1.0	1.0	1.0	1.0	1.0
1.20	1.20	1.0	1.0	1.0	1.0	1.0	1.0

Φ = genetic inheritance (GI) risk level; Π = environmental exposure (EE) risk level; INT = genotype-environment interactions; gene models: Rec = recessive, Dom = dominant, Add = additive.

Note: ^ = INT from equation 6; * = INT from equation 6a; $\alpha \leq 10^{-8}$.

The results shown in Table 7.12 suggest that the GI-EE interactions are very sensitive to low EE levels (Π < 1.10). They also accurately characterize an interaction power value of zero when Π = 1.00, (i.e., no EE risk).

Table 7.13. Main effects (ME) power values, by risk profile (Φ and Π)—all gene models, *N* = 200,000

		Experiment 1			Experiment 3		
Φ	Π	ME^ Rec	ME^ Dom	ME^ Add	ME* Rec	ME* Dom	ME* Add
1.10	1.00	.68	.68	.68	.68	.68	.68
1.10	1.05	.67	.70	.73	.70	.70	.73
1.10	1.10	.65	.75	.68	.75	.75	.68
1.10	1.15	.65	.68	.72	.68	.68	.71
1.10	1.20	.77	**.81**	.69	.78	.79	.68
1.15	1.00	.60	**.88**	**.94**	.60	**.88**	**.94**
1.15	1.05	.65	**.84**	**.92**	.67	**.85**	**.92**
1.15	1.10	.61	**.84**	**.92**	.61	**.84**	**.92**
1.15	1.15	.75	**.92**	**.97**	.75	**.92**	**.96**
1.15	1.20	.65	**.87**	**.94**	.66	**.85**	**.94**
1.20	1.00	.75	**1.0**	**1.0**	.75	**1.0**	**1.0**
1.20	1.05	.73	**1.0**	**1.0**	.75	**1.0**	**1.0**
1.20	1.10	**.81**	**1.0**	**1.0**	**.80**	**1.0**	**1.0**
1.20	1.15	**.81**	**1.0**	**1.0**	**.81**	**1.0**	**1.0**
1.20	1.20	**.85**	**1.0**	**1.0**	**.85**	**1.0**	**1.0**

Φ = genetic inheritance (GI) risk level; Π = environmental exposure (EE) risk level; ME = main effects; gene models: Rec = recessive, Dom = dominant, Add = additive.

Note:^ = ME from equation 7; * = ME from equation 7a; $\alpha \leq 10^{-8}$.

These results suggest that for very large studies it is possible to predict positive associations between recessive genes linked to phenotypes with low to moderate risk.

Conclusions

In summary, the chances of predicting an association in a GWAS are reduced if an environmental effect is present and the statistical model does not adjust for it. This is especially true if the environmental effect and genetic marker do not have an interaction effect. The functional form of the model also matters. The more accurately the form of the environmental influence is portrayed by the

statistical model, the more accurate the prediction will be. Even with very large sample sizes, association predictions involving recessive markers are low.

This study focused on one important methodological step involved in conducting a GWAS: selecting a statistical method and a supporting model that reliably predict associations. This study does not address the broader issue of the supporting experimental design that employs the statistical methods as part of an overall solution strategy. Those combined issues and their mutual interconnections are described by Cordell.[28]

The specific scenarios we address here involve genetic associations that have environmental influences. Our assumption is that the environmental influence that contributes to a given phenotype is in question and the precise form of that influence is unknown. A separate analysis to characterize the functional form to proxy the mechanism behind the environmental exposure is required. These approaches should focus on case-only data similar to the methods described in the Cornelis and colleagues study.[18] These approaches involve investigating different environmentally related functional relationships between the suspected environmental influence and the phenotype in the cases-only subpopulation. For example, if gene effects and environmental effects are independently significant with respect to disease prevalence, a polynomial model could be used to characterize the relationship between environmental effects and the log-odds of disease prevalence. This would allow us to test whether the nonlinear parameterization would be required to characterize the environmental effect. Alternatively, if the environmental effect has multiple levels such as age, researchers could investigate a cubic polynomial to assess whether the effect stayed low initially then rose at some point and flattened out toward the end of the environmental effects range. If this analysis suggests an appropriate polynomial level for environmental effects, researchers should also investigate a similar assessment using the gene-environment interaction variable.

We have used this simulation scenario in previous studies. We reviewed single-gene models and evaluated a wide class of statistical methods.[23] Our results indicated that researchers should consider a multitest procedure that combines individual gene-based (dominant, recessive, additive) core tests as a composite statistical method for conducting the initial screen in a GWAS. The tests can be combined into a single operational test in a number of ways. Two such tests are Holm's Bonferroni procedure and the MAX procedure described by Li and colleagues.[29,30] Of course, if the gene model under investigation is known, a single test that assumes the implied inheritance form is better than

a combined test. For this study, all patterns across gene models are consistent and only vary by degree.

In Chapter 5, we have also evaluated the effect of phenotype errors that resulted from inaccurate diagnoses and genotype errors that resulted from gene-chip errors or occurrences of DNA methylation altering gene expression that associate a wild-type gene with the wrong phenotype outcome.[20] Our results quantify the relationship between genotype and diagnosis error measures and sample size to achieve a .80 statistical power level. Our results also demonstrate that researchers should not underestimate the need to increase sample size to compensate for power loss due to the presence of genotype and diagnosis errors.

In Chapter 6, we also investigated epistatic scenarios involving two genes.[31] The results showed that the most powerful statistical methods for predicting associations between phenotypes and genotypes in epistatic scenarios are statistical models that simultaneously test for associations involving both interacting loci. This is consistent with the results we present here. This result is not surprising and has been reported by others. We reported that if two genes contribute to a phenotype, the weaker gene will be obscured by the stronger gene and often will not be identified as a contributor to the phenotype when a single-gene model is used. Again, this result is similar to showing that the effect of an environmental exposure can obscure the influence of a genotype-phenotype association if the model does not account for the GI and the EE simultaneously. In this sense, two-gene models (or alternatively a gene-environment model) produce better predictions of association than single-gene models do.

We acknowledge that our results could possibly depend on the particular experiments we devised to investigate how the statistical models performed. In light of this, we are reviewing other scenarios to establish the robustness of our findings. Nevertheless, establishing the genotype-to-phenotype connections without using a simulation approach is limited.

For the gene-environment interaction scenarios addressed here, the results across all gene models lead us to conclude that using a composite test that supports distinct underlying statistical models—that is, a "main effects–only" model and a "main effects with interactions" model—is likely to be more effective than single-model tests. This result does not depend on the gene model and thus differs from the single-gene and epistatic scenarios, in which each different gene model assumption (i.e., recessive, dominant, or additive) requires representation in the composite test.[20,29]

Chapter References

1. Cooley PC, Clark RF, Folsom RE. Statistical methods that identify genotype-phenotype associations in the presence of environmental effects. RTI Press Publication No. RR-0022-1405. Research Triangle Park, NC: RTI Press; 2014.
2. Kuo CL, Feingold E. What's the best statistic for a simple test of genetic association in a case-control study? Genet Epidemiol. 2010;34(3):246-253.
3. Suhre K, Shin SY, Petersen AK, et al. Human metabolic individuality in biomedical and pharmaceutical research. Nature. 2011;477(7362):54-60.
4. Spencer C, Hechter E, Vukcevic D, et al. Quantifying the underestimation of relative risks from genome-wide association studies. PLoS Genet. 2011;7(3):e1001337.
5. Dunn OJ. Multiple Comparisons among Means. Journal of the American Statistical Association. 1961;56(293):52-&.
6. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.
7. Terry PD, Umbach DM, Taylor JA. APE1 genotype and risk of bladder cancer: evidence for effect modification by smoking. Int J Cancer. 2006;118(12):3170-3.
8. Stern MC, Johnson LR, Bell DA, et al. XPD codon 751 polymorphism, metabolism genes, smoking, and bladder cancer risk. Cancer Epidemiol Biomarkers Prev. 2002;11(10 Pt 1):1004-11.
9. Browning BL, Browning SR. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. Genet Epidemiol. 2007;31(5):365-75.
10. Zhao J, Jin L, Xiong M. Nonlinear tests for genomewide association studies. Genetics. 2006;174(3):1529-38.
11. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N* Engl J Med. 2000;343(2):78-85.
12. Pearce CL, Rossing MA, Lee AW, et al. Combined and interactive effects of environmental and GWAS-identified risk factors in ovarian cancer. Cancer Epidemiol Biomarkers Prev. 2013;22(5):880-90.

13. Rothman N, Garcia-Closas M, Chatterjee N, et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nat Genet. 2010;42(11):978-84.

14. Lindstrom S, Schumacher F, Siddiq A, et al. Characterizing associations and SNP-environment interactions for GWAS-identified prostate cancer risk markers--results from BPC3. PLoS One. 2011;6(2):e17142.

15. Yu K, Wacholder S, Wheeler W, et al. A flexible Bayesian model for studying gene-environment interaction. PLoS Genet. 2012;8(1):e1002482.

16. Patel CJ, Chen R, Kodama K, et al. Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus. Hum Genet. 2013;132(5):495-508.

17. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. Am J Epidemiol. 2009;169(2):219-26.

18. Cornelis MC, Tchetgen EJ, Liang L, et al. Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. Am J Epidemiol. 2012;175(3):191-202.

19. Kraft P, Yen YC, Stram DO, et al. Exploiting gene-environment interaction to detect genetic associations. Hum Hered. 2007;63(2):111-9.

20. Cooley P, Clark RF, Page G. The influence of errors inherent in genome wide association studies (GWAS) in relation to single gene models. J Proteomics Bioinform. 2011;4:138-144.

21. Schymick JC, Scholz SW, Fung HC, et al. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol. 2007;6(4):322-8.

22. Iles MM. Effect of mode of inheritance when calculating the power of a transmission/disequilibrium test study. Hum Hered. 2002;53(3):153-7.

23. Cooley P, Clark R, Folsom R, et al. Genetic inheritance and genome wide association statistical test performance. J Proteomics Bioinform. 2010;3(12):321-325.

24. Chan EK, Hawken R, Reverter A. The combined effect of SNP-marker and phenotype attributes in genome-wide association studies. Anim Genet. 2009;40(2):149-56.

25. Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007;445(7130):881-5.

26. Ziegler A, Konig IR, Thompson JR. Biostatistical aspects of genome-wide association studies. Biom J. 2008;50(1):8-28.

27. Lehmann EL, Romano JP. Testing statistical hypotheses. 3rd ed. New York: Springer; 2005.

28. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet. 2009;10(6):392-404.

29. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6:65-70.

30. Li Q, Zheng G, Li Z, et al. Efficient approximation of P-value of the maximum of correlated tests, with applications to genome-wide association studies. Ann Hum Genet. 2008;72(Pt 3):397-406.

31. Cooley P, Gaddis N, Folsom R, et al. Conducting genome-wide association studies: epistasis scenarios. J Proteomics Bioinform. 2012;5(10):245-251.

CHAPTER 8

Polygene Methods in Genome-Wide Association Studies (GWAS)

Philip Chester Cooley and Ralph E. Folsom

Overview

The hope that the Human Genome Project would pave the way to detect associations between genetic markers and common diseases such as heart disease, diabetes, auto-immune diseases, and psychiatric disorders has not achieved expectations.[1] Genome-wide association studies (GWAS) have been the dominant tool used in the exploration of the genome to determine the associations between genomic regions and complex traits in the population. There are two ways to improve the GWAS process. The first is to improve the quality of the data; the next generation of sequence data will provide greater genomic coverage and will soon be available to the scientific community. The second is by improving the computational process. Issues such as multiple testing and the influence of external (nondisease) variables that confound experimental results limit the interpretability of the statistical results.
This chapter examines the utility of a computation process that considers multiple markers that simultaneously associate with a phenotypic trait. This computational process builds a single nucleotide polymorphism (SNP) network using a stepwise association method that proceeds in stages adding a single SNP at each stage until no further additions are called for.

Initially, the model used in a GWAS consisted of a series of single-locus statistical tests, examining each SNP independently for association to the phenotype. In Chapter 4, we examined this single-gene model and showed that, if the mode of inheritance (MOI) model—whether genes are dominant, recessive, additive, or multiplicative—is unknown, then a Cochran-Armitage composite (CA-C) test, which pools the possibility of three MOI outcomes into

This chapter has not been published elsewhere.

a single test, is the most powerful single-gene model test. We also showed that the classical case-control method of epidemiology (an MOI-agnostic test) will never be as powerful as the CA-C test.

Because most GWAS assume that the phenotype and the genotype data are error free, we examined the consequences of this assumption in Chapter 5 and found that adherence to the assumption forecasts sample size requirements too small to achieve reliable association predictions and undermines replication across studies.

Because genes do not always act as single triggers but act in concert with other genes, polygene analysis is an important methodology that represents the next step forward if GWAS are to remain a viable exploratory tool. Single-gene GWAS have detected associations with numerous diseases but explain little about disease heritability. Polygenic models assume that thousands of genetic variants could impact the phenotype. For example, in humans, height, skin color, eye color, and weight are all phenotypes linked to polygenes.

Chapter 6 discusses epistatic effects in association studies. We describe a simple polygene method: two-gene models that have been designed to act in ways that depend on the pair of MOIs and how the pair are positioned relative to each other. We show that in contrast to the single-gene model, two-gene models that are MOI agnostic (i.e., the two-gene case-control model) are more powerful than either single-gene models or models that treat gene-gene interactions only.

Chapter 7 uses a similar approach as that discussed in Chapter 6, but we replace one of the genes with a variable representing an external environmental influence such as a chemical exposure or the effects of aging. We show that detection of the genetic marker is more likely to occur if the external process is represented in the statistical model even if the characterization is inaccurate.

This chapter also builds on the work of previous chapters by confronting the issue of how to process general polygene models with an unknown number of significant traits. We propose stepwise polygene analysis as a new strategy for studying the association between large sets of SNP predictors and groups of correlated phenotypes (i.e., outcomes). These data commonly arise in GWAS, in which associations with a large number of qualitative and quantitative phenotypes are investigated using hundreds of thousands of genetic markers. Our strategy uses the log linear-logistic model framework suited to analyzing qualitative trait responses. In the two-gene epistatic scenario, we showed that single-locus models do not detect all of the markers that are part of

the phenotype pathway. Other studies have also reported this finding.[2] However, from a combinatorial perspective, polygene models can create an overwhelming number of comparisons when using more than two loci. A number of methods exist that have developed approaches to examine all-possible two-locus combinations.[3,4] However, the computational requirements of these approaches is daunting.

Chapter 8 defines a new polygene approach: a multistep process that is analogous to a stepwise association process. A major new contribution is that the initial stage of our process examines all usable SNP autosomes using the CA-C test—which, as we showed in Chapter 4, is the most powerful of the conventional tests used to predict associations in GWAS[5]—to predict each SNP's genetic inheritance properties. The accuracy of these inheritance predictions determine the effectiveness of the method because this assumption is used to estimate the type of wild-type alleles for each loci. Subsequent stages statistically combine the SNPs from previous stages and examine whether any of the unassigned SNPs are associated with the phenotype based on the SNPs from previous stages. Because each stage builds on the previous one, we refer to this method as the "stepwise association method," and we use it to predict polygene networks for a given phenotype.

To test the method, we generated sets of simulated data described in the "Generating SNP Data" section. These data sets consist of SNP (genotype) data for each subject and an associated phenotype (case-control) measure. The interclass correlation of the genotype data is controlled to determine how strongly the different SNPs resemble each other, and each SNP has a designated MOI and risk level. Using a combination of SNP and genotype information, we estimate the value that each SNP contributes to the probability of a case and use Monte Carlo methods to assign a case-control measure. Collectively, this constitutes a "truth set" of known outcomes and the factors that affect genotype-phenotype association.[6]

Hence, we can assess the performance of each method to predict known outcomes. In general, our results suggest that the process is effective for main effect and interaction SNPs with modest effects, as measured by their odds ratios (ORs), but variables with small effects present a challenge that will be difficult to overcome even with very large sample sizes.

Background

In the past 7 years, GWAS have been the most widely used tool in genetic research. Yet over this period, the tools of our research efforts have failed to unravel the complete biological architecture of genetic diseases.[7] Researchers have learned from the myriad GWAS conducted during the past few years that the biggest effect sizes for associations between genes and phenotypes are much smaller than anticipated. For example, in a genome-wide meta-analysis of intelligence quotient (IQ) involving nearly 18,000 children, the largest effect size accounted for only 0.2 percent of the variance.[8] This suggests that, in general, smaller effects will be barely detectable in GWAS and extremely difficult to replicate. However, this finding conflicts with hundreds of candidate gene and gene-environment interaction studies that find significant effects using modest sample sizes. The finding also implies that many of the reported markers are false positives that cannot be replicated, although some journals now require that candidate gene papers include an independent replication.[9]

Identifying additional loci of small effect can be partially accomplished by meta-analysis of multiple GWAS using stringent significance thresholds. Furthermore, it is unlikely that GWAS will ever be powered to identify the full spectrum of small effects. Several recent analytical approaches have been developed to test whether common variants of extremely small effect size could be combined to explain phenotype variation. This approach was successfully applied to the schizophrenia gene *SCZ*. A polygene small-effect SNP model explained approximately 35 percent of disease diagnosis variance of an estimated total of 80 percent.[10] In a second example, Yang and colleagues estimated that 67 percent of the heritability of human height could be explained by a polygenic model.[11] Although both models fail to estimate total heritability, the models do account for far more than that of known, validated associations. However, these approaches are limited because they are unable to identify the proportion that each marker contributes to trait variation: the marginal effects contributed by each variable are impossible to estimate using models with many genetic variables.[12] Nevertheless, a polygenic analysis defines a large set of variants with an unknown subset that affects phenotype. Together, these represent the true underlying biology.

Performing polygenic analysis to understand the genetic basis of complex traits leads to a systems biology perspective in which many perturbations of a complex network contribute to the outcome of a complex trait phenotype. The complexity is difficult or impossible to disentangle on a per variant basis. To

establish the biology of the complex trait directly, researchers need a systems genetics approach in which large sets of genetic variants and/or genes are analyzed in an integrated way and incorporate functional data.[13]

Disease-Scoring Methods

Disease-scoring methods are one such approach that researchers have explored. Combining multiple genetic markers into a single score that predicts disease risk is a relatively new approach for associating SNPs with disease in the context of GWAS. This approach has shown that some diseases have a strong genetic basis, even if few actual genes have been identified. Disease scoring has also revealed a common genetic basis for distinct diseases. The scoring approach has been used to obtain evidence of genetic effects when no single markers are significant, establish a common genetic basis for related disorders, and construct risk prediction models. Published studies have demonstrated that significant associations of polygenic scores have occurred only in well-powered studies and that useful levels of prediction may occur when predictors are estimated from very large samples, up to an order of magnitude greater than any currently available. This suggests the need for studies to use larger sample sizes.[14]

Many geneticists believe that better polygenic scoring methods will be developed. Quantitative genetic techniques that estimate heritability using DNA suggest that about half of the heritability can be detected using the common SNPs that are currently genotyped on commercially available DNA arrays, given sufficiently large samples.[15] Economic improvements in whole-genome sequencing will make it cost-effective to identify DNA sequence variation of every kind throughout the genome.[7]

Studying environment effects is more difficult than studying genes because, as Cooley and colleagues have shown, the environment is more complex than DNA with its genotype 0,1,2 designation code.[16] However, environmental research likely could capitalize on the advances in whole-genome technology to identify biomarkers of environmental influence and therefore help detect genetic associations.[7]

Statistical Approaches to Polygene Analysis

Although disease scoring is one viable approach to polygene analysis, this chapter focuses on using statistical analyses to approach the problem. In an earlier study of epistatic loci, we showed that for a given locus, single-locus tests are not as effective as two-locus tests for predicting associations if the risk

value for a second interacting locus exceeds a relative risk of 1.05–1.12.[1] The crossover risk value varies, depending on the genetic inheritance properties of the pair of interacting loci. In general, the power of two-locus tests to detect associations improves as the risk value of the second locus increases, whereas the power of single-locus tests progressively declines.

For certain inheritance models and risk values, a true association between a locus and phenotype can be masked by a second interacting locus when using single-locus tests. However, these findings are not unexpected and are consistent with the findings of others.[17-19] We also note that a study of ALS subjects identified SNPs that, when paired with other SNPs, became significant markers of ALS, although there were no indications of association using single-gene models.[20] Thus, some markers can only be identified if they are paired with other markers within a polygene network that is part of the phenotype's biological pathway.*

Multilocus analyses are not as straightforward as conducting single-locus tests and present numerous computational, statistical, and logistical challenges.[21] Examining all pairwise combinations of SNPs using gene chips that generate 500,000 to 1,000,000 SNPs is currently computationally infeasible. One approach to this issue is to preprocess the SNPs' set to eliminate redundant SNPs. A way to accomplish this is to select a set of results from a single-SNP analysis based on an arbitrary significance threshold and exhaustively evaluate interactions in that subset. We incorporate this approach into our stepwise polygene analysis method but acknowledge that selecting SNPs to analyze based on main effects will prevent certain multilocus models from being detected where the heritability is concentrated in the interaction rather than in the main effects.

Our stepwise polygene analysis approach looks for a main effect at Stage 1 but also considers the possibility of interactions. The benefit of our method is that it performs an unbiased analysis for interactions within the selected set of SNPs. It is also far more computationally and statistically tractable than analyzing all possible combinations of markers.

Bush and colleagues describe another strategy: restricting examination of SNP combinations to those that fall within an established biological context,

* We use the term "biological pathway" to represent the biological reactions and the interaction network in a cell where each reaction is identified with its enzyme, which in turn is coded by certain genes. Some of the most common biological pathways are involved in metabolism, regulating gene expression, and transmitting signals. Pathway analysis plays a key role in the advanced studies of genomics.

such as a biochemical pathway or a protein family.[22] Because both these techniques rely on electronic repositories of structured biomedical knowledge, they use a tool that generates SNP-SNP combinations with a statistical method that evaluates combinations in the GWAS data set. For example, the Biofilter approach uses public data sources in conjunction with logistic regression and multifactor dimensionality reduction methods.[23] Similarly, the INTERSNP tool for genome-wide interaction analysis (GWIA) of case-control SNP data and quantitative traits selects SNPs for joint analysis using a priori information. Sources of information to define meaningful strategies are based on statistical evidence (e.g., single-marker association studies computed from different data sources). INTERSNP uses logistic regression and contingency table approaches to assess SNP-SNP interaction models.[24]

Cooley and colleagues examined the performance of three different two-locus tests.[25] In most scenarios, the two-locus, case-control 8-degree of freedom (*df*) Pearson test was the most powerful. In certain scenarios (i.e., when both genes have a dominant MOI), the unlinked, cases-only version of the Wu and colleagues test as refined by Ueki and colleagues was optimal. The version of the Ueki test in which cases and controls are included was never optimal.[26,27] These findings are not surprising because Wu and colleagues' modified test measures interaction effects exclusively, whereas the two-locus, case-control test includes main effects for both loci as well as interaction effects.

One unresolved issue is how to construct a computationally practical test that takes into account interactions and enhances the detection of associations between a specific locus and the phenotype of interest. As mentioned previously, Wang and colleagues conducted an empirical comparison of five epistatic interaction detection methods, which examines all combinations of two-gene methods.[3] Each of the five methods demonstrated unique utilities, but no single method was simultaneously the most powerful and the most scalable and had the lowest type-1 error rate in every setting. When users want powerful results and are not concerned with computation cost, Wang and colleagues cite the TEAM method of Zhang and colleagues as having the highest performing algorithm.[3,4] However, researchers should note that even limiting the number of interacting genes to two requires $n \times (n - 1) / 2$ association calculations. For n = 500,000–1,000,000, the computational requirements are daunting but readily parallelizable.

Methods

The Stepwise Algorithm

Our method proceeds in stages. At each stage, we add an unassigned SNP to the network unless the test score is below the significance threshold. The first step (Stage 1) makes an initial pass against all usable SNP autosomes to identify the most highly significant SNP-phenotype association. The second stage statistically combines the Stage 1 SNP with all original autosome SNPs to identify the most significant SNP pair associated with the phenotype. This stage uses a test for significance that is conditional on the Stage 1 SNP. We then continue this process for triple SNPs, quadruple SNPs, and so on until combining the loci produces no new SNP-phenotype associations. This process is analogous to a stepwise regression process, in which networks of SNPs are connected stage by stage until no new SNPs can be identified. Our process is not an exhaustive polygene analysis, but it does assume that at least one loci associated with the phenotype can be identified via a single-gene model.

Step 1 Details. The initial pass uses the CA-C test to examine all SNP autosomes and apply a restrictive significance threshold to identify the most highly significant SNP-phenotype association, known as the Stage 1 SNP.

The single-locus model used in Stage 1 is based on the CA-C test that used an MOI agnostic model that combined the three versions (recessive, dominant, and additive) of the classical CA-C method into a single test. Our previous results indicate a multitest procedure that combines the results of individual MOI-based tests is an effective method for initially filtering a GWAS. Note that the CA-C method can be used to predict the SNP-specific MOIs. Consequently, the first stage not only selects the Stage 1 SNP, it also tags all other SNPs with an MOI, which is then used to predict ensuing stage-specific SNPs.

Ensuing Steps. Stage i (i = 2,3,4, …) is defined as progressively adding one SNP, if appropriate, into the process; the stages advance until it is no longer possible to add SNPs and advance. Stage i applies a model of all of the Stage i-1 SNPs and computes a test score. Stage i (i > 0), i-1 SNPs have been associated with the phenotype and are assumed to be part of the SNP/phenotype network. We then systematically add remaining unassigned SNPs to the set of stage i-1 SNPs and compute a new test score. The test score is defined as twice the difference between the natural log of the likelihood including the Stage i

SNP and the natural log of the likelihood including only the Stage i-1 SNPs. This test score has approximately a χ^2 distribution with one degree of freedom if the log odds ratio for the Stage i SNP is actually 0.

Note that the test used to identify the Stage i (i>1) SNP uses a logistic regression (LR) approach and the construction of the test depends on the stage. A log likelihood (LLH) criteria is used to estimate the LR parameters and measure whether the Stage i SNP sufficiently differentiates between cases and controls given the set of Stage i-1 SNPs at a specified significance level T.

Formally, for s_i to be a candidate SNP at Stage i, it must satisfy:

$$t_i = \text{LLH}[s_i / s_1 , s_2 , \ldots, s_{i-1}] \geq T \text{ for } i = 2,3,4, \ldots N, \tag{8.1}$$

where $s_1 \neq s_2 \neq \ldots \neq s_{i-1}$. If more than one SNP satisfies equation (8.1), s_i is defined as the SNP with the largest value of t_i. If $t_i < T$ the process is terminated at the i-1 stage. Also note that T will be Bonferroni corrected at each stage in the process to account for the multiple number of tests carried out for each stage.

To summarize, the process begins with the best single SNP and continues for SNP pairs, triple SNPs, quadruple SNPs, and so forth, until the process is unable to identify any new SNPs to add to the network. At Stage i, a test of significance is used that is conditional on the prior selected set of SNPs. The process continues to define a networks of SNPs that are connected stage by stage, until no new SNPs can be identified.

Generating SNP Data

To assess the method we propose, we first had to generate a synthetic set of SNP data. There are a number of features that characterize these data that also characterize real-life properties of SNP data:

1. The SNP inheritance properties
2. The distribution of genotype alleles
3. The correlation between SNP pairs (as a linkage disequilibrium measure)
4. The contributions of the individual SNPs toward their association with the phenotype

We discuss each of these properties in subsequent sections. Note that alternative methods to generate data are presented in the Appendix to this chapter.

Mode of Inheritance. There are two main properties of genetic inheritance formulated by Mendel: the principle of segregation and the principle of independent assortment. In segregation, for any particular trait, the pair of alleles of each parent separate and only one allele passes from each parent on to an offspring. In independent assortment, different pairs of alleles are passed to offspring independently of each other. This principle indicates why the human inheritance of a particular eye color does not increase or decrease the likelihood of having six fingers on each hand, for example. This is because the genes for independently assorted traits are located on different chromosomes. The Online Mendelian Inheritance in Man (OMIM)[28] provides the best source of information on the MOI distribution (see Table 8.1). However, OMIM is disproportionally populated by genes linked to single Mendelian disorders, and genes associated with multifactorial disorders are underrepresented. Because polygene influences are assumed to be a major source of additive and multiplicative SNP behavior, the distribution in Table 8.1 is likely biased.

Table 8.1. Distribution of genes in Online Mendelian Inheritance in Man, by mode of inheritance (MOI)

MOI	Frequency
Autosomal Dominant	3,805
Autosomal Additive	12
Autosomal Multiplicative	21
Autosomal Recessive	3,775

Genotype Distributions. The genotype distribution data we used in this study was first described by Schymick et al.[6] It consists of 555,352 SNPs from 276 ALS patients and 271 neurologically normal controls. These data are publicly available to the scientific community and were produced by the Laboratory of Neurogenetics of the intramural program of the National Institute on Aging (NIA), National Institutes of Health (NIH). The genotyping was performed using the Illumina Infinium assay humanhap550. Infinium assays assess haplotype tagging SNPs based upon Phases I+II of the International HapMap Project.

Linkage Disequilibrium (LD). LD is a property of SNP pairs that describes the degree to which an allele of one SNP is inherited (correlated) with an allele of another SNP. The concept of LD was developed by population geneticists in an attempt to describe changes in genetic variation within a population.

According to the theory, recombination events occur within a family. This effect is amplified through generations, and in a population of fixed size undergoing random mating, repeated random recombination events will eventually produce linkage equilibrium/independence. Linkage between markers on a population scale is referred to as "linkage disequilibrium."

In other words, LD is the nonrandom association of alleles at two or more loci that descend from single, ancestral chromosomes. It is also a measurement of distance between SNP locations and represents the combination of alleles in a population more often or less often than would be expected from a random draw. The degree of LD depends on the difference between observed allelic frequencies and those expected from a homogenous, randomly distributed model.[29] Linkage equilibrium occurs in populations in which combinations of alleles or genotypes can be found in the expected proportions.[30]

However, biology is not the only factor at play with respect to LD measurements. Humans tend to associate with other humans who have similar genetic characteristics. Studies have shown that populations tend to stratify according to nonrandom mating patterns in which subjects with similar genotypes and/or phenotypes mate more frequently than would be expected under a random mating pattern. For example, it is common for individuals of similar body size to mate. Evidence for this process has been reported from the National Longitudinal Study of Adolescent Health.[31] One feature of this study tested genetic similarity between friends. Using friendship networks clusters of genotypes demonstrated genotype positive correlation (homophily) and negative correlation (heterophily). These results were replicated in an independent sample from the Framingham Heart Study. These unique results suggest that association tests should take into account the fact that human genes are influenced by the humans of their social networks.

A natural way to measure LD is by the correlation coefficient.[29] Characterizing LD structure is useful for designing genetic association studies because we can control for genetically homogeneous populations and their influence on population-specific parameters. However, in reality, a study sample can be genetically heterogeneous with substructure even if the recruiting criterion requires a specific ethnic group.

In this study, we specify the degree of correlation (using the product moment or Pearson correlation coefficient) between each SNP pair during the generation of the SNP database as a measure of LD.

Designating the Case-Control Assignments. The final procedure in creating the SNP data set uses the SNP data (with the appropriate LD genomic spacing) and the SNP odds ratio (ODDS) data to determine the case (1) control (0) designation codes that are used to predict association. In Table 8.2, we present three examples of data we will use to test our method.

We assume that these three examples are real gene networks that are known with certainty. Our goal is to examine the genotype data that describe these networks to assess method performance. To accomplish this, we will create data sets that reproduce the specifications that define the networks. These specifications include the MOI, which we limit to dominant, recessive, and additive; sample size (N); the LD characteristics; and the ODDS relative to the penetrance odds. If P0 is the disease penetrance probability, then the penetrance odds is P0/(1-P0)=exp(B0). If P1 is the probability of disease when a given dominant MOI SNP (SNP1) has one or two wild-type alleles, then P1/(1-P1)=exp(B0+B1) is the disease odds with the addition of SNP1. The SNP1 ODDS divided by the penetrance odds or the SNP-1 odds ratio (relative to the penetrance odds) is therefore exp(B1).

By knowing these specifications, we can reverse-engineer a data set that replicates (with varying degrees of certainty) the starting specifications. This approach thus generates replicates with varying degrees of goodness of fit. We specify the total number of replicates to calculate and then store the best-fitting replicates as the data to analyze, because their LR coefficients best conform to the specifications and serve to represent the method's average performance over a large number of replicated samples of the specified size.

Example 1 consists of 7 high-risk SNPs (ODDS > 1.2), 4 low-risk SNPs (ODDS > 1.0 & ODDS < 1.2), and 4 no-risk SNPs (ODDS = 1.0). The penetrance is .25. There are 5 recessive SNPs, and 10 dominant SNPs.

Example 2 is identical to Example 1, except it includes two additional effects that both represent interactions. The first is the interaction between SNPs 3 and 6. The second represents the interaction between SNPs 1 and 9. All four SNPs that compromise the interaction terms have positive main effects.

Example 3 comprises three SNPs with ODDS = 1.10, three no-risk SNPs, and nine SNPs with risks between 1.01 and 1.05. This example is designed to test the feasibility of performing association studies with low-effect SNPS, and thus will require a much larger sample size to detect an effect.

In summary, the X (design) matrix consists of genotype data that contains an LD structure specified exogenously. We use these data, in conjunction with

the MOI and the ODDS parameters, to determine the probability of a case. Using a random number compared with the case probability, we assign the case-control designation vector Z and the number of subjects N, which we also specify exogenously and use to determine the number of rows in X and Z, as Table 8.2 shows.

Table 8.2. SNP data set design data—three examples

Variable	MOI	Parameter	Example 1 Value	$E^{(value)}$ = ODDS	Example 2 Value	$E^{(value)}$ = ODDS	Example 3 Value	$E^{(value)}$ = ODDS
Intercept		B_0#	-1.386	.25	-1.386	.25	-0.916	.40
SNP1	D	B_1	.2231	1.25	.2231	1.25	.0010	1.01
SNP2	R	B_2	.0	1.00	.0	1.00	.0953	1.10
3	D	B_3	.3365	1.40	.3365	1.40	.0953	1.10
4	D	B_4	.0953	1.10	.0953	1.10	.0010	1.01
5	R	B_5	.4055	1.50	.4055	1.50	.0	1.00
6	D	B_6	.1823	1.20	.1823	1.20	.0488	1.05
7	D	B_7	.1397	1.15	.1397	1.15	.0392	1.04
8	R	B_8	.0	1.00	.0	1.00	.0953	1.10
9	R	B_9	.2852	1.33	.2852	1.33	.0488	1.05
10	R	B_{10}	.0953	1.10	.0953	1.10	.0198	1.02
11	R	B_{11}	.2231	1.25	.2231	1.25	.0010	1.01
12	D	B_{12}	.3001	1.35	.3001	1.35	.0392	1.04
13	D	B_{13}	.0953	1.10	.0953	1.10	.0	1.00
14	R	B_{14}	.0	1.00	.0	1.00	.0	1.00
15	D	B_{15}	.0	1.00	.0	1.00	.0296	1.03
SNP1 × SNP9	I	B_{16}	X	X	.2523	1.30	X	X
SNP3 × SNP4	I	B_{17}	X	X	.1823	1.20	X	X

MOI = mode of inheritance; ODDS = SNP odds ratios; SNP= single nucleotide polymorphism.

#B_0 = log odds of penetrance.

We assign a constant genomic distance between SNP pairs as measured by the Pearson correlation coefficient, R. We begin by selecting SNP1 at random from the Schymick et al.[6] distribution of SNPs. We define D as the genomic distance, which will be reflected in R. We then select D percent of the SNP1 genotypes and assign them to SNP2; the remaining 1-D percent are selected at random. This

process performs the same calculations with SNP2 and SNP3, and so forth. Thus, the genomic distance between SNP1 and SNP2 will be approximately D and the genomic distance between SNP1 and SNP3 will be D squared.

Completion of this step generates a design matrix, X, consisting of genotype values (0,1,2). The next step, which calculates the case-control designation proceeds according to the following steps:

1. Given that the X_{ij}s are = {0,1,2}F, define a set of variables Y_{ij}s derived from the X_{ij}s in the following way:
   a. If $X_{ij} = 2$ and j is a recessive SNP, $Y_{ij} = 1$ otherwise $Y_{ij} = 0$.
   b. If $X_{ij} > 0$ and j is a dominant SNP, $Y_{ij} = 1$ otherwise $Y_{ij} = 0$.
   c. Ignore (for the time being) additive and multiplicative SNPs.
2. Calculate the case score $W_i = B_0 + SUM\ (B_j \times Y_{ij})$ for the ith subject.
3. Convert the score into a probability $p_i = \exp(W_i) / (1.0+\exp(W_i)$ for the ith subject (p_i is the predicted probability of a case).
4. Use p_i to generate Z_i .the designation of the case-control value i.e. if random number (0–1) < p_i $Z_i = 1$ otherwise $Z_i = 0$.
5. After the Z vector is generated, determine whether the calculation of the B_js are sufficiently close to the parameters in Table 8.2 that were used to generate the data set by performing a logistic regression of the generated data to see how well the estimated coefficients reproduce the values in Table 8.2.
6. Generate a measure of "fit closeness" with respect to the estimated B_is relative to the parameters in Table 8.2 and save the corresponding design matrix with the closeness measures. This will generate a number of replicates that exhibit closeness scores, with the last one having the best overall correspondence.

At this point in the process we have generated an X matrix and a companion Y matrix, which indicates whether the genotype has a wild-type allele. We also generated a vector Z that accounts for the collective ODDS of the SNPS. Furthermore, the process has been checked by estimating the coefficients (i.e., step 5 above) of the logistic regression model for a number of replicates. We save only the replicates that match original ODDS specifications for future analysis, which in our case will be to assess how the proposed method performs.

Results

Model Comparisons and MOI Predictions

The first example illustrates the approach using genotype data consisting of SNP pairs that are paired at rates D = 0.0, 0.2, and 0.4. We assume that the 15 SNPs have risk levels that are consistent with those that Table 8.2 presents. We then use two LR models to apply our method. The first model uses the 0,1,2 genotype codes as the independent variable codes. The second model uses the MOI predictions to stratify the genotype data into wild-type and non–wild-type allele. We will refer to each step as a stage.

The Stage 1 MOI predictions are shown in Table 8.3. The missed predictions predominate for those SNPs with low ODDS values and for LD = .4. For example, with ODDS > 1.0, the missed predictions occur twice in the LD = .0 column (SNPs 10 and 13), once in the LD = .2 column (SNP 6) and three times in the LD = .4 column (SNPs 5, 7, 11).

Table 8.3. Stage 1 summary

SNP	ODDS	MOI—Actual	MOI—Predicted LD = .0	MOI—Predicted LD = .2	MOI—Predicted LD = .4
1	1.25	D	D	D	D
2	1.00	R	R	D	D
3	1.40	D	D	D	D
4	1.10	D	D	D	D
5	1.50	R	R	R	**A**
6	1.20	D	D	**A**	D
7	1.15	D	D	D	**A**
8	1.00	R	D	A	A
9	1.33	R	R	R	R
10	1.10	R	**A**	R	R
11	1.25	R	R	R	**A**
12	1.35	D	D	D	D
13	1.10	D	**A**	D	D
14	1.00	R	D	A	D
15	1.00	D	R	D	A

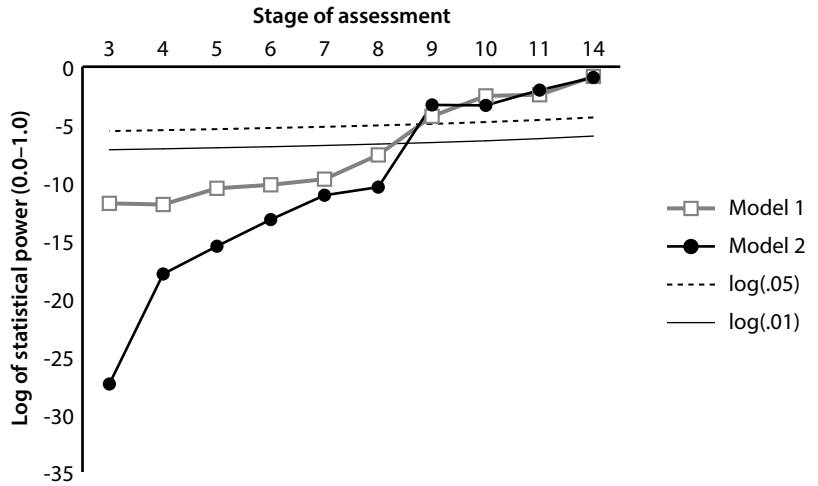
SNP = single nucleotide polymorphism; ODDS = SNP odds ratios; MOI = mode of inheritance; LD = linkage disequilibrium; *N* = 10,000.

Note: Missed predictions are shown in boldface.

Having predicted MOIs reasonably accurately, the question remains: is there value in using this information to predict associations? Figures 8.1A, 8.1B, and 8.1C compare the two models for each LD level. The three figures display the log of the test *p*-value scores by stage. The log values of the corrected .05 and .01 threshold levels are also included on each figure. They make a clear case that Model 2, which uses the MOI information to map the three genotype values to the two allele levels, is more powerful than Model 1, which treats the genotype levels as a single variable. Both models use the LR framework. In each of the figures, the *p*-value of the Model 2 (orange) is less than the *p*-value of Model 1 (blue).

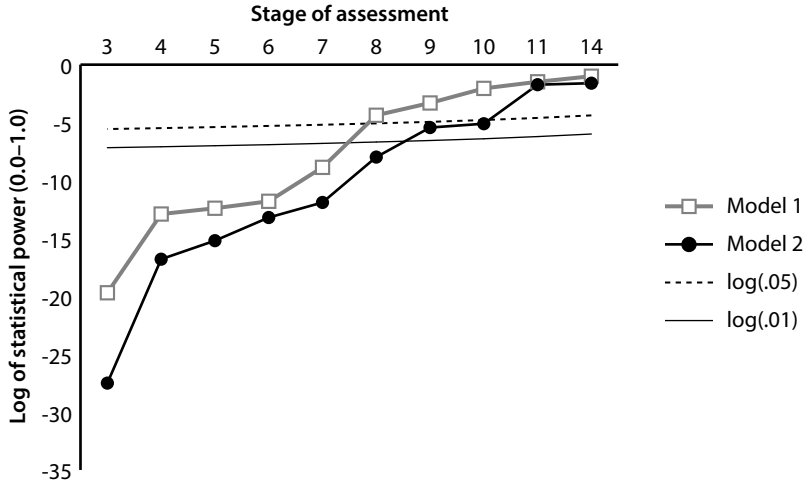
Table 8.4 translates the *p*-value measures in the figures into a performance measure describing the composition of the network. This table summarizes the overall performance by model for each LD level. For example, the comparison for LD = .2 indicates that at the .01 (.05) threshold, Model 1 missed 6 (3) SNPs (of 11 SNPs with ODDS > 1.0) and associated 5 (8) correctly. By comparison, Model 2 predicted correctly 10 (9) SNPs at the .01 (.05) significance level and missed 1 (2).

Figure 8.1a. Model comparison for LD = 0.0 SNP data



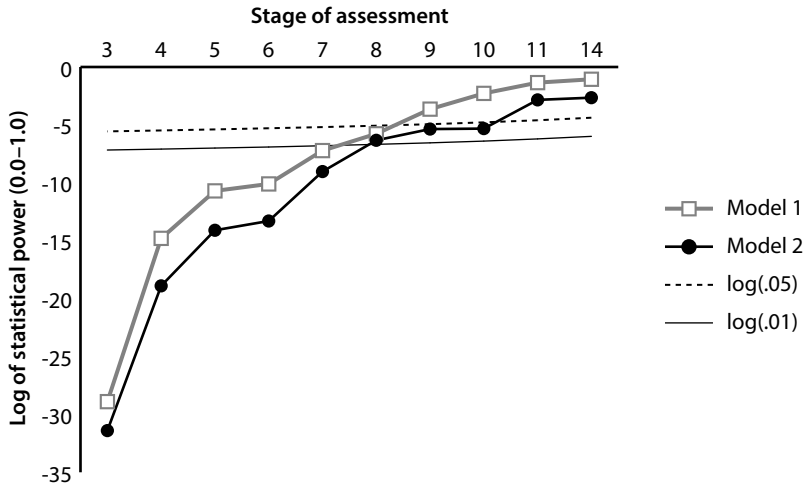
LD = linkage disequilibrium; SNP = single nucleotide polymorphism; MOI = mode of inheritance; Model 1 = model without predicted MOI; Model 2 = model with predicted MOI; log(.05) = .05 corrected confidence threshold; log(.01) = .01 corrected confidence threshold.

Figure 8.1b. Model comparison for LD = 0.2 SNP data



LD = linkage disequilibrium; SNP = single nucleotide polymorphism; MOI = mode of inheritance; Model 1 = model without predicted MOI; Model 2 = model with predicted MOI; log(.05) = .05 corrected confidence threshold; log(.01) = .01 corrected confidence threshold.

Figure 8.1c. Model performance comparison for LD = 0.4 SNP data



LD = linkage disequilibrium; SNP = single nucleotide polymorphism; MOI = mode of inheritance; Model 1 = model without predicted MOI; Model 2 = model with predicted MOI; log(.05) = .05 corrected confidence threshold; log(.01) = .01 corrected confidence threshold.

Table 8.4. Prediction accuracy results, by model and LD criteria

Criteria\LD	.0	.0	.2	.2	.4	.4
Model	1	2	1	2	1	2
Correct at .01	8	8	5	9	7	7
Correct at .05	8	8	8	10	8	10
Missed at .01	3	3	6	2	4	4
Missed at .05	3	3	3	1	3	0
ODDs ≤ .01 missed at .01	4	4	3	3	2	1
ODDS ≤ .01 missed at .05	3	3	2	0	2	0

LD = linkage disequilibrium; ODDS = SNP odds ratios; SNP = single nucleotide polymorphism; *N* = 10,000.

The results of Table 8.4 and Figures 8.1A through 8.1C suggest collectively that differences exist between statistical models and that Model 2 outperforms Model 1. Also, with the subject size *N* at 10,000, the predictive accuracy of both models appears unreliable for those SNPs with an effect size measured by ODDS ≤ 1.10.

Models With Interactions

Our investigation of interaction effects used Example 2, as defined in Table 8.2. Example 2 is the same as Example 1, with two additional interactions terms. SNP1 and SNP9 form an interaction with ODDS = 1.30, and SNP3 and SNP form an interaction with ODDS = 1.2. SNP9 is recessive; therefore, we would anticipate that this interaction term would be more difficult to detect than the interaction involving SNPs 3 and 4, which are both dominant SNPs.

The first task is to build a data set that conforms to the Example 2 specifications. We do this for each of the three SNP intercorrelation effects that varied from .0 to .2 to .4. Table 8.5 shows the build results.

The results shown in Table 8.5 are based on assigning *N* = 10,000, running 1,000 replicates, and selecting the run that best fits the specification from the 1,000 replicates. In reconstructing the data, fitting the specifications is not an exact process and even small deviations could influence the accuracy of the association prediction phase. However, we controlled this by adding replicates and we generated sufficient replicates to demonstrate that the method's performance is consistently applied. The LD = .0 and .2 runs do not identify one main effect SNP (with an ODDS < .10) and both interactions (type II errors). For the LD = .4 run, the MOI of SNP 5 was inaccurately predicted as an additive SNP.

Table 8.5. Build results for interaction example (Example 2) by LD levels = 0.0, 0.2, and 0.4

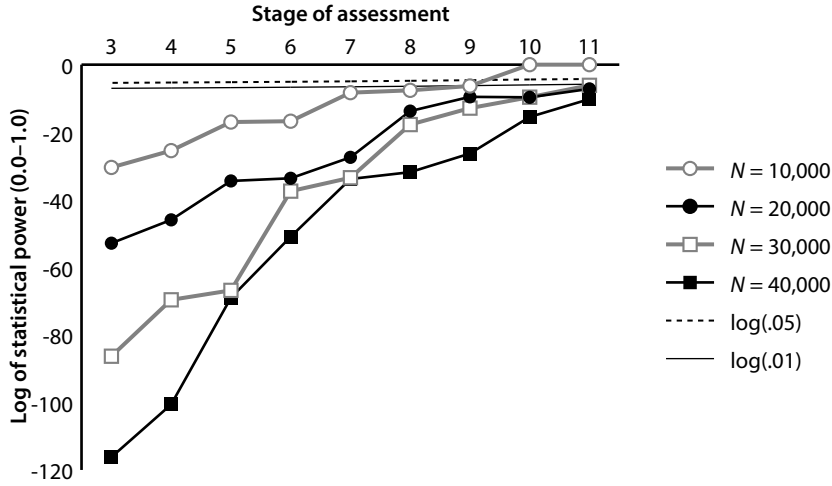
SNP	ODDS	MOI	Corr =.0 ODDS reconstructed	Corr = .2 ODDS reconstructed	Corr = .4 ODDS reconstructed
1	1.25	D	1.26	1.29	1.23
2	1.00	R	.96	.99	1.02
3	1.40	D	1.41	1.42	1.40
4	1.10	D	1.13	1.16	1.03
5	1.50	R	1.48	1.55	1.52
6	1.20	D	1.24	1.24	1.19
7	1.15	D	1.15	1.15	1.16
8	1.00	R	1.02	.98	.97
9	1.33	R	1.33	1.35	1.36
10	1.10	R	1.10	1.11	1.09
11	1.25	R	1.20	1.25	1.30
12	1.35	D	1.34	1.37	1.38
13	1.10	D	1.12	1.08	1.11
14	1.00	R	.97	1.00	.97
15	1.00	D	1.00	1.02	1.03
1 × 9	1.30	D × R	1.31	1.27	1.26
3 × 4	1.20	D × D	1.18	1.24	1.19
Correct Calls			10	10	8
Missed			SNP 10 & Interactions	SNP 13 & Interactions	SNPs 5,11,13 & Interactions

SNP = single nucleotide polymorphism; ODDS = SNP odds ratios; MOI = mode of inheritance.

Overall, these errors contributed to the exclusion of SNP5 as a member of the network. The central question is "can the interactions be predicted for larger sample sizes?" Figure 8.2 results partially address this issue by graphing the log of the *p*-value scores by algorithm stages as a function of *N* (i.e., *N* = 10,000, 20,000, 30,000 ,and 40,000).

Table 8.6 summarizes the results of Figure 8.2 for *N* = 40,000 subjects. An *N* = 40,000 is sufficient to predict all SNPs with ODDS > 1.00. Only one of the two interactions was predicted to associate with the phenotype. The interaction involving recessive SNP9 was undetected.

Figure 8.2. Influence on sample size *N* on model performance



LD = linkage disequilibrium; SNP = single nucleotide polymorphism.

Table 8.6. Results by stage: correlated SNP pairs LD = .200

Stage	ODDS	MOI	SNP	LD = .20 *p*-value
1	1.40	D	3	-269.2
2	1.35	D	12	-132.3
3	1.33	R	9	-116.2
4	1.50	R	5	-100.6
5	1.25	D	1	-69.02
6	1.20	D	6	-50.95
7	1.10	D	4	-33.88
8	1.25	R	11	-31.83
9	1.15	D	7	-26.31
10	1.10	D	13	-15.49
11	1.20	D × D	3 × 4	-10.20
12	1.10	R	10	-6.135

ODDS = SNP odds ratios; MOI = mode of inheritance; SNP = single nucleotide polymorphism; LD = linkage disequilibrium; *N* = 40,000, note all significant at the .01 (Bonferroni-corrected) level. Only the SNP1 by SNP9 interaction was not detected. All SNPs with ODDs = 1.00 were not associated (correctly).

The results of Figure 8.2 and Table 8.6 suggest that more than $N = 10{,}000$ samples are needed before the interaction is detected. However, one of the interaction terms involving the recessive SNP (SNP9) was undetected even with an $N = 40{,}000$.

Detecting Low-Effect SNPs

Table 8.7 presents results using Example 2. The results are intended to examine the sample-size requirement for Model 2 with SNP data that has low levels of risk. Recall that the Example 2 network consists of three SNPs with ODDS = 1.10, three no-risk SNPs, and nine SNPs with risks between 1.01 and 1.05. We expanded our sample to 300,000 to detect the low-level effects contained in the example.

Table 8.7. Results of low-effect genes by stage and correlated SNP pairs LD = .200 for Model 2

Stage	ODDS	MOI	SNP	*N* = 100,000 log(*p*-value)	ODDS	MOI	SNP	*N* = 200,000 log(*p*-value)	ODDS	MOI	SNP	*N* = 400,000 log(*p*-value)
1	1.10	D	3	-43.45	1.10	D	3	-59.77	1.10	D	3	-115.0
2	1.05	D	2	-18.16	1.04	D	7	-31.03	1.10	R	8	-75.33
3	1.05	R	7	-15.39	1.10	R	8	-24.94	1.10	R	2	-43.84
4	1.10	R	8	-10.77	1.10	R	2	-21.03	1.05	D	6	-37.24
5	1.10	R	6	-8.504	1.05	R	9	-13.72	1.04	D	12	-24.97
6	-	-	-	-	1.05	D	6	-12.98	1.05	R	9	-23.39
7	-	-	-	-	1.04	D	12	-11.35	1.03	D	15	-10.42
8	-	-	-	-	1.03	D	15	-7.835	1.04	D	7	-9.722
9	-	-	-	-	-	-	-	-	1.01	D	4	-6.457
Missed	1.04	D	4	-5.854	1.01	R	11	-5.101	1.02	R	10	-1.740
""	1.03	D	9	-3.677	1.02	R	10	-1.740	1.01	D	1	-1.274
""	1.04	D	12	-3.614	1.01	D	1	-1.274	1.00	R	5	-.8376
""	1.02	R	15	-2.412	1.00	R	5	-.8376	1.00	D	13	-.80
""	1.01	D	1	-2.046	1.00	D	13	-.8067	1.01	D	4	-.5392
""	1.00	D	10	-1.729	1.01	D	4	-.5392	1.00	R	14	-.1960
""	1.01	R	14	-1.488	1.00	R	14	-.1960	-	-	-	-
""	1.01	D	11	-1.246	-	-	-	-	-	-	-	-
""	1.00	R	13	-1.062	-	-	-	-	-	-	-	-
""	1.00	R	5	-.3823	-	-	-	-	-	-	-	-

ODDS = SNP odds ratios; MOI = mode of inheritance; SNP = single nucleotide polymorphism.

As Table 8.7 indicates, increasing *N* affects the performance of the stepwise algorithm. For *N* = 100,000, the algorithm identifies five significant loci at the .01 corrected significance level. For *N* = 200,000, eight loci are identified, and for *N* = 400,000, nine significant loci are identified. Overall, the method missed seven small-effect loci when *N* = 100,000, four loci when *N* = 200,000, and three loci when *N* = 400,000.

Table 8.7 also demonstrates that a very large *N* is necessary to detect very low risk levels (<1.03), especially if the SNP is recessive. Also, much of the *N* = 400,000 model advantage occurs at its higher stages, which suggests that even models with larger Ns may still be ineffective for low-risk loci.

Overall, these results suggest there is little value in extending *N* to levels higher than 400,000.

Discussion

This study has demonstrated that creating a set of synthetic genes with known properties that can be analyzed in the context of GWAS experiments is possible and informative. Because the properties of the genes are determined, and it is possible to establish exploratory protocols that define computational best practices. At least seven other similar approaches have been reported in the literature; however, the creation process we used in these experiments controls for LD, sample size, MOI, the single locus main effect risks, and the interaction loci risks, which other similar approaches do not.

Our synthetic gene procedure is designed to examine the properties of defining a network of genes that link to a single phenotype. The method for establishing this gene network is straightforward. It proceeds in stages and uses logistic regression methods at each stage except Stage 1. Each stage uses a maximum likelihood ratio test to identify the optimal locus to include in the network, given the loci that have accumulated in the network in prior stages. The Stage 1 calculation comprises the composite test that uses the three variations of the Cochran-Armitage test. Each variation assumes one of three MOIs and permits an estimate of the individual loci MOI. The MOI prediction is then used in subsequent stages to improve statistical power.

The results indicate that

- a network of SNPs that links to a given phenotype can be identified if (and only if) the study size is sufficient;
- interaction effects can be addressed in this process but may require a larger *N* in their detection;

- the statistical power of the method may be sensitive to the level of LD; and
- a network consisting of very low-effect loci can be predicted accurately if the study size, N, is sufficient, but the study size needed is far larger than that employed by traditional GWAS. For example, to detect very low-effect SNPs, we used $N = 400{,}000$. This is well beyond what is conceived as practical, but future research and improvements might make it possible.

Many polygene methods seek to investigate a large number of SNP combinations. These approaches are computationally demanding and often impractical to implement. Even with widespread recognition that single-locus tests are likely to be inferior to multilocus tests for GWAS of many diseases and phenotypes, an unresolved issue is how to construct a computationally practical test that takes into account interactions and enhances the detection of associations between multiple loci and the phenotype of interest.

The main value of our study is that it proposes computations that are practical and straightforward and require less computational effort than those reviewed by others.[3] Furthermore, our simulation studies suggest that if the associations are real, they can be found with properly powered studies. An additional feature of our study is that as part of the Stage 1 SNP selection, the prediction defines the inheritance properties of all SNPs, which our simulation experiments have shown to be accurate and can be used to improve the reliability of ensuing stage predictions.

Appendix: DNA Simulator Tools

A number of different coalescent simulators that generate simulated DNA fragments for different evolutionary models are discussed in the literature that differ from the approach we describe in the main text. These applications simulate different standard neutral evolutionary models with recombination, variable population size, and migration. They also allow for spatial and temporal environmental heterogeneity. Carvajal-Rodríguez provide a comprehensive description of many of these models.[32] The most recent and one of the most versatile models is not included in his analysis. GWAsimulator is a program that simulates genotype data for SNP chips that are used in GWAS.[33] It creates whole genome case-control or population samples. It also can simulate specific genomic regions. For case-control data, the program can be linked to sample cases and controls according to a user-specified multilocus disease model. The program requires phased data as input, and the simulated data will have similar LD patterns as the input data.

A second program worth mentioning is GENOME.[34] It simulates a wider range of scenarios including recombination hotspots. As well as whole genome data. In addition to features of standard coalescent simulators, the program allows for recombination rates to vary along the genome and for flexible population histories. The program and C++ source code are available online at http://www.sph.umich.edu/csg/liang/genome/.

Chapter References

1. Visscher PM, Brown MA, McCarthy MI, et al. Five years of GWAS discovery. Am J Hum Genet. 2012;90(1):7-24.
2. Chatterjee N, Wheeler B, Sampson J, et al. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet. 2013;45(4):400-5, 405e1-3.
3. Wang Y, Liu G, Feng M, et al. An empirical comparison of several recent epistatic interaction detection methods. Bioinformatics. 2011;27(21):2936-43.
4. Zhang X, Huang S, Zou F, et al. TEAM: efficient two-locus epistasis tests in human genome-wide association study. Bioinformatics. 2010;26(12):i217-27.
5. Cooley P, Clark R, Folsom R, et al. Genetic inheritance and genome wide association statistical test performance. J Proteomics Bioinform. 2010;3(12):321-325.
6. Schymick JC, Scholz SW, Fung HC, et al. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol. 2007;6(4):322-8.
7. Plomin R, Simpson MA. The future of genomics for developmentalists. Dev Psychopathol. 2013;25(4 Pt 2):1263-78.
8. Benyamin B, Pourcain B, Davis OS, et al. Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. Mol Psychiatry. 2014;19(2):253-8.
9. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011;187(2):367-83.

10. International Schizophrenia Consortium, Purcell SM, Wray NR, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460(7256):748-52.

11. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565-9.

12. Gibson G. Hints of hidden heritability in GWAS. Nat Genet. 2010;42(7):558-60.

13. Mackay TF, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. Nat Rev Genet. 2009;10(8):565-77.

14. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013;9(3):e1003348.

15. Plomin R, Haworth CM, Meaburn EL, et al. Common DNA markers can account for more than half of the genetic influence on cognitive abilities. Psychol Sci. 2013;24(4):562-8.

16. Cooley PC, Clark RF, Folsom RE. Statistical methods that identify genotype-phenotype associations in the presence of environmental effects. RTI Press Publication No. RR-0022-1405. Research Triangle Park, NC: RTI Press; 2014.

17. Culverhouse R, Suarez BK, Lin J, et al. A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet. 2002;70(2):461-71.

18. Hoh J, Wille A, Zee R, et al. Selecting SNPs in two-stage analysis of disease association data: a model-free approach. Ann Hum Genet. 2000;64(Pt 5):413-7.

19. Li J, Horstman B, Chen Y. Detecting epistatic effects in association studies at a genomic level based on an ensemble approach. Bioinformatics. 2011;27(13):i222-9.

20. Sha Q, Zhang Z, Schymick JC, et al. Genome-wide association reveals three SNPs associated with sporadic amyotrophic lateral sclerosis through a two-locus analysis. BMC Med Genet. 2009;10:86.

21. Moore JH, Ritchie MD. STUDENTJAMA. The challenges of whole-genome approaches to common diseases. JAMA. 2004;291(13):1642-3.

22. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012;8(12):e1002822.

23. Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pac Symp Biocomput. 2009:368-79.

24. Herold C, Steffens M, Brockschmidt FF, et al. INTERSNP: genome-wide interaction analysis guided by a priori information. Bioinformatics. 2009;25(24):3275-81.

25. Cooley P, Gaddis N, Folsom R, et al. Conducting genome-wide association studies: epistasis scenarios. J Proteomics Bioinform. 2012;5(10):245-251.

26. Wu X, Dong H, Luo L, et al. A novel statistic for genome-wide interaction analysis. PLoS Genet. 2010;6(9):e1001131.

27. Ueki M, Cordell HJ. Improved statistics for genome-wide interaction analysis. PLoS Genet. 2012;8(4):e1002625.

28. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine. Online Medelian Inheritance in Man (OMIM). 2016 [cited 2016 Feb 11]; Available from: http://www.ncbi.nlm.nih.gov/omim

29. Lin CY, Xing G, Xing C. Measuring linkage disequilibrium by the partial correlation coefficient. Heredity (Edinb). 2012;109(6):401-2.

30. Abecasis GR, Noguchi E, Heinzmann A, et al. Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet. 2001;68(1):191-197.

31. *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. Add Health: The National Longitudinal Study of Adolescent ot Adult Health. 2015 [cited 2015 July 20]; Available from: http://www.cpc.unc.edu/projects/addhealth

32. Carvajal-Rodriguez A. Simulation of genomes: a review. Curr Genomics. 2008;9(3):155-9.

33. Li C, Li M. GWAsimulator: a rapid whole-genome simulation program. Bioinformatics. 2008;24(1):140-2.

34. Liang L, Zollner S, Abecasis GR. GENOME: a rapid coalescent-based whole genome simulator. Bioinformatics. 2007;23(12):1565-7.

CHAPTER 9

Conclusions and Recommendations

Philip Chester Cooley

Purpose of the Manuscript

Our explorations into genome-wide association studies (GWAS) led us to believe that a number of governing principles were absent from the manner in which GWAS were conducted. The descriptions of methods that researchers have provided in the literature commonly assumed an additive inheritance model or were agnostic with respect to an inheritance model assumption. Furthermore, there were many examples in which GWAS provided inconsistent results and researchers could not replicate study findings. Our own experiences investigating GWAS indicated that different statistical methods assuming specific inheritance properties also produced different association results. Because the inheritance properties are generally not known a priori, we favored a statistical model that was agnostic with respect to inheritance, and we turned to simulation studies to confirm our beliefs.

We chose simulation because we noted the absence of both methodological standards and a process for evaluating statistical methods used in predicting associations between genes and phenotypes. We believed that that we could examine these missing elements more effectively with a simulation approach. Accordingly, we created a simulated data set of virtual genes that were linked to known outcomes. These data constituted a "truth set." We analyzed the simulated data using different statistical methods and used it to assess each method's predictive properties. This process revealed a number of principles described subsequently.

Summary of the Methodology

We developed a database of synthetic/virtual genetic loci. We assigned Mendelian inheritance properties to those loci and connected different combinations/networks of loci to virtual phenotypes. The phenotypes were qualitative traits intended to represent the positive or negative diagnoses of specific diseases. We also introduced errors into the recording of the genotypes and the assignment of diagnoses, and we simulated the strength of the genotype to phenotype signal via a relative risk process. For each inheritance, error, and relative-risk setting combination, we assigned a targeted number of cases and controls to include in an experimental GWAS design. The number of cases was varied in an effort to encapsulate a statistical power region containing the value of 0.8.

Having generated these data, we then applied a number of distinct GWAS statistical methods and models that were germane to the specific GWAS, and we used them to predict an association. We replicated each unique GWAS design by statistical prediction model 1,000 times.

Each database entry, including performance measures, was catalogued with respect to the individual data entries and the properties that we used to generate GWAS data. An analysis of these measures across the replicates permitted performance assessment of the different statistical methods to predict genotype-to-phenotype associations.

Our experiments were not limited to genetic mechanisms. We also generated database entries that incorporated environmental influences into the GWAS experimental designs that were used to simulate different mechanisms such as aging and chemical exposure.

We also investigated loci that depended on the properties of other loci to produce phenotypic elevated risk levels. The effect of one gene that depends on the presence of one or more "modifier genes" is known as epistasis, and others have shown that epistatic mutations have different effects in combination than individually because of interactions between genes or within them that lead to nonadditive effects.[1]

By design, every entry contained in the database was produced with a positive genetic association risk. The risks were varied between 1.01 and 1.50 and were kept small by design to test the sensitivity of the statistical models. To assess method performance, we applied many different and varied statistical methods to determine which class of models exhibited the best power properties. Because all entries had a positive risk, each and every method

should have predicted an association. Statistical power was determined by computing the percentage of positive predictions across the 1,000 replicates. Graphing the power profiles as a function of GWAS subject participants enabled us to estimate the number of subjects required to achieve a specific power goal and also illustrated how errors in the data, signal strength, environment, and polygene behaviors affected model predictions.

Findings and Recommendations

The analysis described previously has led us to formulate the following assertions:

- Developing a synthetic gene database that recorded known outcomes between synthetic phenotypes and genotype networks provides a mechanism that was not possible using real genomics data. Of course it has been asserted that real data and data generated by simulation experiments may not have the same qualities (i.e., real data will contain more noise than simulated data). Our response was to provide a mechanism that generates data that includes a level of noise that can be specified a priori.
- The creation of this database enabled an evaluation of different statistical models and methods specifically because the prediction outcomes were known and statistical power profiles could be estimated.
- Methods that treat combinations of genes acting in concert either with other genes or environmental factors are important investigations in a GWAS context because a single-gene model will fail to identify markers in many types of gene-gene, and gene-environment networks.
- In response to this principle, we proposed a general polygene test (Chapter 8) that is based on a procedure that accepts (or not) a positive association of the n+1 SNP given the presence of a set of *n* previously confirmed SNPs. In the absence of a test that considers the combined effect of multiple genes, the association process will be dominated/obscured by the strongest (dominant) associated locus.
- A reliable single-gene model is still necessary to identify a starting point/filter in a polygene process, but the influence of the inheritance properties of the locus should be considered in the selection of the statistical model. There can be a substantial power loss if the inheritance property of the loci is assumed to be additive (or log-odds additive), and the locus is dominant (which results in a modest power loss) or recessive (which results in a

substantial power loss). Of course, if the gene model is additive, a model that assumes additive behavior is optimal with respect to power loss.

- Making an informed assessment of the inheritance properties of the locus is possible. Applying multiple tests that each assume a distinct inheritance property and selecting the property that produces the lowest *p*-value is one possibility. Knowing the inheritance will improve the performance of the polygene process.
- If we assume error-free phenotype data, we may need to supplement the number of case/control subjects to achieve desired power targets (substantially, if the locus is recessive). Adding noise to both phenotype and genotype measures is informative in this context.
- The GWAS process is exploratory in nature, and various tools, such as odds ratio confidence regions and pseudo R-squared (R^2) measures, are useful to the exploration. However, odds ratio confidence regions and pseudo R^2 are limited to logistic regression model explorations.
- It is possible to use the public databases that link SNPs to genes to diseases/conditions and to develop interesting and plausible explanations of genotype-phenotype linkages. In our experience, computing cell counts often exposes potential biases due to low frequency of disease alleles.

Future Directions

GWAS have identified many new genetic risk factors for a number of common human diseases, but much work remains to be done and can only be accomplished by using new approaches. Furthermore, new technologies are coming, such as whole-genome sequencing, which will replace the 1 million SNP chip data with the entire genomic sequence of 3 billion nucleotides.

This will have a huge impact on data storage, manipulation, quality control, and data analysis processes. New computer science and bioinformatics tools will be needed, and cloud operations will become the new infrastructure. Also, high-throughput technologies for measuring the transcriptome, the proteome, the environment, and the whole-genome sequences will become standard operating procedures. New phenotypes from the development of new technologies such as neuroimaging will also become available and add to the level of complexity.

In this environment, unravelling genotype-phenotype relationships for the purpose of improving health care will expand. Integrating these varied and complex biomedical data and findings is the future of human genetics.

Chapter References

1. Otto SP, Feldman MW. Deleterious mutations, variable epistatic interactions, and the evolution of recombination. Theor Popul Biol. 1997;51(2):134-47.

Acknowledgment

The development of this book was made possible by the generous support of the RTI Fellow program.

Contributors

Philip Cooley, MSc, is a senior fellow at RTI International in Computational Biology and High Performance Computing. He has more than 47 years of experience in developing computer models used in the study of environmental health and disease transmission scenarios. He has developed infectious disease models to study the transmission of a number of diseases including malaria, tuberculosis, HIV/AIDS, and influenza. He is currently developing a process to forecast stunting in Indonesia, and his research focus is shifting to the development and application of models that explore and predict the impending epidemic of chronic diseases.

Robert F. Clark, PhD, is a senior genetic epidemiologist in RTI International's Genetic Epidemiology and Omics research program. Throughout most of his career, he has focused on multidisciplinary work in the omics and systems biology of various multifactorial disorders, and since 1992, he has conducted many genetic studies of neurodegenerative diseases; breast, bone, and brain cancers; nicotine, heroin, and cocaine addiction; and aortic aneurysms.

Ralph E. Folsom, PhD, chief scientist in RTI's Division of Statistical and Data Sciences, is an expert in the design and analysis of complex probability samples. Working on the nation's largest household survey (the National Survey on Drug Use and Health or NSDUH), Dr. Folsom initiated innovative weight adjustment methods based on his logistic response propensity and exponential poststratification models. This pioneering work led to the sophisticated GEM weight adjustment methods currently employed for NSDUH. Dr. Folsom also introduced model-based imputations for missing frequency of use and income data items, and he has been an influential collaborator in the development of NSDUH's current Predictive Mean Neighborhoods (PMN) imputation methodology. Dr. Folsom has recently led RTI's innovative work in small area estimation research. In addition to his innovative work on many complex survey efforts, Dr. Folsom has made significant contributions to the development of RTI's computer software for survey data analysis, SUDAAN.

Nathan Gaddis, PhD, is a research programmer/analyst in RTI's Research Computing Division. He has a diverse background in molecular biology and bioinformatics, combining laboratory research experience in microbiology, immunology, and genetics with genome-scale analyses, including genetic analyses for GWAS and transcriptome-wide analyses of alternative splicing. Currently, Dr. Gaddis is a co-investigator on several GWAS projects, the lead developer for the National Heart, Lung, and Blood Institute-funded LungMAP project, and a developer for the National Human Genome Research Institute-funded PhenX project.

Grier Page, PhD, is a senior statistical geneticist in RTI's Genomics and Statistical Genetics Research unit. He has been conducting research in all areas of statistical genetics since 1993 and in systems biology since 2002. Dr. Page has developed methods for the analysis of linkage and association data as well as microarray, proteomic, and next-generation sequencing methods. Dr. Page has also developed numerous bioinformatics tools to implement the methods he has developed, including the PowerAtlas (http://www.PowerAtlas.org), HDBStat! (http://www.ssg.uab.edu/HDBStat), and CressExpress (http://www.cressexpress.org).

Diane Wagener, PhD, was an epidemiologist at RTI International. She has 37 years of experience in academia, government, and consulting studying the causes, genetics, and social impact on a number of diseases, both in the United States and internationally. Her expertise is in research design and statistical and computational analysis of the data.

Index

Note: Page numbers followed by *t* and *f* indicate tables and figures.

E

F

G

M

T

U

V

W

Y

Z

This groundbreaking work uses a simulated data set to evaluate new analytic methods in genome-wide association studies (GWAS). The human genome is very complex, and the effect of a genetic variant depends on many factors including where the gene is expressed, when it is expressed, how it interacts with other genes that themselves may harbor variants, and the effect of the environment. GWAS have identified many new genetic risk factors for a number of common human diseases, but much work remains to be done and can only be accomplished by using new approaches. The role of this book is to help jump-start the investigation of such new approaches. Identifying which computational strategy is best suited for investigating a specific aspect of genomics is a daunting task. Using simulated data to evaluate new analytic methods provides a "truth set" against which to assess methods' predictive properties. In this book, Cooley and colleagues use simulated data to test a variety of analytic methods, starting with single-gene models and progressing to more complex polygene and gene-by-environment scenarios. The methods that Cooley and colleagues use are straightforward, easily applied, and thoroughly documented.

RTI Press

RTI INTERNATIONAL

www.rti.org/rtipress

RTI Press publication BK-0016-1608