NOVEL STATISTICAL METHODS FOR THE STUDY DESIGN AND ANALYSIS OF GENOME-WIDE ASSOCIATION STUDIES

by
Lindsey Allen Ho

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics, Gillings School of Global Public Health.

Chapel Hill
2010

Approved by:

Dr. Ethan M. Lange, Advisor
Dr. Fred A. Wright, Committee Member
Dr. Fei Zou, Committee Member
Dr. Yun Li, Committee Member
Dr. Kari E. North, Committee Member

ABSTRACT
LINDSEY ALLEN HO: NOVEL STATISTICAL METHODS FOR THE STUDY DESIGN AND ANALYSIS OF GENOME-WIDE ASSOCIATION STUDIES.
(Under the direction of Dr. Ethan M. Lange.)

In Chapter 2, we compare the power of association studies using cases and screened controls to studies that incorporate free public control genotype data. We describe a two-stage replication-based design, which uses free public control genome-wide genotype data in the first stage, and follow-up genotype data on study controls in the second stage. We assess the impact of systematic ancestry differences and batch genotype effects. We show that the proposed two-stage replication-based design can dramatically increase statistical power and decrease cost of large-scale genetic association studies.

In Chapter 3, we describe and compare conventional haplotype analysis approaches to a number of haplotype sharing measures. We evaluate the impact of the inclusion of markers in linkage disequilibrium (LD) on power and assess the utility of recoding scores using thresholds. Finally, we develop a quick and novel approach based on categorizing similar haplotypes into contingency tables. These alternative methods are compared via simulation assuming a rare-recessive disorder caused by a small number of high-penetrant mutations within a single disease locus. We found that incorporating allele frequencies and dichotomizing scores increased power. Conversely, using fixed windows and excluding single nucleotide polymorphisms (SNPs) in low LD or with low minor allele frequencies decreased power. Finally we show that our novel clustering algorithm had competitive power than permutation testing.

In Chapter 4, we describe an alternative method to single SNP analyses of single or multiple candidate genes that is designed to increase power when multiple SNPs are associated with the trait. Our method is based on forward selection in regression that

provides a joint test of the statistical significance of a gene. Within the framework of a simulated candidate gene study as well as a study of related candidate genes, we assess the power of this method by simulating a quantitative trait and compare our proposed method to single SNP and other multiple SNP models. Our results suggest that our method is competitive to conventional methods and may be more powerful when SNP x SNP interactions exist.

DEDICATION

My dissertation is a reflection of the steadfast love and inspiration of primarily my dear wife, spiritual partner, and closest friend, Elena, as well as of my children, Camilo (nine years old) and Mayeli (six years old). They have daily offered their unconditional encouragement and understanding. For them, I reserve a profound and beloved place in my heart for empowering me to attain this achievement. I also dedicate my dissertation to my parents, Lloyd and Lena Ho, for instilling in me at an early and tender age the desire to reach the highest, promoting the completion of my degrees, and assisting my family and me throughout the years. Last but not least, I convey my sincere appreciation to my sister, Liana, and her family, Jason, Malia, and Jayden for their unwaivering support and compassion, in addition to my brother, Langley, and his family, Chloe and Preston.

On a light-hearted yet genuine note, I express my gratitude to Faji, my faithful fuzzy friend, pet, and cat, who lovingly accompanied me throughout numerous hours of my studies and research.

ACKNOWLEDGMENTS

I earnestly thank my dissertation chair, Dr. Ethan Lange, for his guidance and support. Ethan continually urged me to produce work of the highest quality and thoroughness. He encouraged me to astutely assess my results and reasoning at all stages of the scientific process. I appreciate the countless hours he devoted to our discussions that aided in deepening my understanding of statistical genetics. I also thank Dr. Leslie Lange for always offering a light-hearted disposition to us whenever we were in the thickets of our investigations, along with Mr. Yunfei Wang for his persistent and contagious cheerfulness when working together in our laboratory.

I express my gratitude to my committee members, Drs. Fred Wright, Fei Zou, Kari North, and Yun Li, for their input, suggestions, critical comments, and interest. It was a pleasure to work with them.

I recognize the financial support I received from the National Institute of Environmental Health Sciences Predoctoral Training Grant in Environmental Biostatistics (NIEHS grant T32 ES007018), which assisted in providing a solid academic and practical training in biostatistics and also served as a springboard to conduct independent research for my dissertation in addition to preparing me for a leadership position in biostatistics. I additionally acknowledge the financial support through my Graduate Research Assistantship under Ethan (start-up funds through the UNC Department of Genetics and National Institutes of Health grants CA120082 and CA1363621), which included complete coverage to participate in conferences.

I am indebted to my colleagues at SRA International, Drs. Rich Cohn, Pat Crockett, Ruchir Shah, Deepak Mav, Marjo Smith, Jessica Matthews, Mike Easterling, Ms. Dhiral Phadke, Ms. Laura Betz, and others, who offered a positive outlet and environment for applied statistics and collaborative research. In particular, Pat has been an incredible role model and leader for me, who has shaped me through our research projects and open exchanges.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

A brief and directed review of seminal analytical techniques that have aided researchers in revealing genetic determinants of disease is contained herein, as well as study design and statistical issues surrounding these approaches. Linkage analysis is discussed in Section 1.2, association studies are reviewed in Section 1.3, and a concise and pertinent list of terms and definitions used in the field of genetics is presented in Section 1.4.

1.2 Linkage analysis

Linkage analysis has traditionally been used in determining disease genes, which involves proposing a genetic model *a priori* that explains disease inheritance and subsequently observe disease status and marker genotype patterns in pedigrees (Lander and Schork, 1994; Ellsworth and Manolio, 1999c). Linkage analysis has particularly been successful with monogenic Mendelian diseases (Jimenez-Sanchez et al., 2001; Hirschhorn and Daly, 2005). Pritchard (2001) and Reich and Lander (2001) have

observed that these monogenic diseases are often due to rare variants (Hirschhorn and Daly, 2005). The mapping resolution of linkage analyses has been reported to be no more than about 1 centiMorgan, on average 1,000 kilobases (Boehnke, 1994).

In contrast to simple Mendelian disorders, common diseases may be the result of the total effect of and/or interactions among multiple genetic and environmental factors (Hirschhorn and Daly, 2005; Wang et al., 2005). As a result, any particular causal gene could have an overall modest effect on disease risk. Furthermore, there have been several reports that linkage analysis is not well powered to uncover these common genetic variants (Risch and Merikangas, 1996; Risch, 2000; Cardon and Bell, 2001; Tabor et al., 2002).

1.3 Association studies

Association studies are a strategic complement to linkage studies and, unlike linkage studies, are powered to identify common genetic variants underlying complex diseases (Risch and Merikangas, 1996; Cardon and Bell, 2001; Tabor et al., 2002; Carlson et al., 2004; Hirschhorn and Daly, 2005). One way to analyze data from an association study is to compare unrelated diseased cases and controls with respect to their frequencies of alleles or genotypes at a given marker. It is also possible to use family-based controls as a way to control for population stratification (Hirschhorn and Daly, 2005).

1.3.1 Population stratification

Population stratification is the existence of multiple subgroups within a population such that the disease prevalence within each subgroup is different (Hirschhorn and Daly, 2005). In association studies, population stratification can result in the overrepresentation of one or more subgroups in the sampled disease cases. This can result

in a false-positive test at a given genetic marker if the allele frequencies differ in the different subgroups. However, there has been a considerable amount of debate as to the extent to which population stratification results in false-positives (Cardon and Palmer, 2003; Marchini et al., 2004; Clayton et al., 2005; Berger et al., 2006).

1.3.2 Family-based sampling designs

To address the possible complications arising from population stratification, family-based sampling designs have been developed to choose the optimal control population. In the parent-parent-affected offspring trio design, genotype data is collected on affected individuals and their parents, and information is used on the alleles transmitted and not transmitted from the parents to the affected offspring. The non-transmitted alleles constitute the control sample, whereas in the unrelated case-control study design, individuals form the control sample. Several methods have been developed to analyze such data (Falk and Rubinstein, 1987; Terwilliger and Ott, 1992; Spielman et al., 1993; Spielman and Ewens, 1996), though it has been noted that Spielman and Ewens' (1996) transmission disequilibrium test (TDT), which has been widely used and is essentially McNemar's test of symmetry for paired data (McNemar, 1947), is subject to technical artifacts due to laboratory difficulties (Mitchell et al., 2003; Hirschhorn and Daly, 2005). In addition, the parent-parent-affected offspring trio design may be biased toward ascertaining younger patients for late onset diseases (Hirschhorn and Daly, 2005).

To circumvent this ascertainment bias for late onset diseases, the discordant sib design was developed, which is a family-based association approach that matches an affected individual with one or more unaffected siblings (Boehnke and Langefeld, 1998; Horvath and Laird, 1998; Spielman and Ewens, 1998). In addition to being immune to population stratification, the discordant sib design allows the control of shared envi-

ronmental effects, assuming that siblings within a family shared the same environment. However, despite the advantages, Horvath and Laird (1998), Morton and Collins (1998), and Spielman and Ewens (1998) have reported that family-based association designs only involving siblings are not as powerful as case-control studies.

1.3.3 Population-based sampling design: unrelated cases and controls

The unrelated case-control study design, a traditional epidemiological tool, has been easy and convenient for studying the relationship between putative genetic risk factors and disease outcome (Schork et al., 2001), though the ease and convenience must be considered in light of its caveats. Population based samples of a large number of affected individuals (i.e. individuals with the disease or trait, cases) and a large number of presumably well-matched unaffected individuals (i.e. individuals without the disease or trait, controls) are collected and one way of evaluating statistical significance is to examine the difference in observed frequencies of the cases' and controls' exposure to the genetic risk factor. If significantly more cases than controls are exposed to the genetic risk factor, then one may deduce that the genetic risk factor is involved in disease pathogenesis, or protective to disease pathogenesis if significantly more controls than cases are exposed.

The main assumption of genetic case-control studies is that the alleles at the locus in question have a causal relationship with disease status (Schork and Chakravarti, 1996). It is further assumed that the genotyped alleles are either at the disease locus or in linkage disequilibrium (LD) with the causal genetic variant. If the alleles are in linkage equilibrium with the disease locus, then the causality assumption is questionable (Schork and Chakravarti, 1996).

Unlike linkage studies that follow inheritance patterns of disease status and geno-

typed markers within pedigrees, the unrelated case-control study design is population based and relationships are unknown, i.e. in the case of haplotypes, information is not captured on the evolution of haplotypes throughout time from the most recent common ancestor. Furthermore, affection status is not followed over generations and analysis rests upon correlations of current disease status with current genotypes (Balding, 2006).

1.3.4 Types of population association studies

Balding (2006) described four main types of population association studies, though a particular study may exhibit characteristics from more than one type. Candidate polymorphism studies investigate a particular polymorphism which is believed to be involved in disease onset. In candidate gene studies, 5 to 50 single nucleotide polymorphisms (SNPs) within a gene are genotyped, the candidate gene being determined from the results of a previous linkage or association study or from prior biological knowledge about the gene's function. Fine mapping studies often probe a candidate region of 1 to 10 megabases with hundreds of SNPs genotyped and possibly spanning 5 to 50 genes. Similar to candidate gene studies, an earlier linkage or association study may have located the candidate region, though unlike candidate gene studies, fine mapping covers a much wider region.

Genome-wide association (GWA) studies are the fourth type of population association study. This approach searches the majority of the genome for genetic variants that give rise to disease. Whereas candidate gene studies can be thought of as a hypothesis-testing approach since positional or functional knowledge motivates these studies, GWA studies represent a hypothesis-generating approach since the genomic location of disease susceptibility variants is not assumed, but rather the aim is to uncover these variants (Borecki and Suarez, 2001; Hirschhorn and Daly, 2005).

Previously, GWA studies were not reasonable to conduct because of the required

extensive labor and high cost (Hirschhorn and Daly, 2005). However, the decreasing cost per genotype coupled with the improving technical ability to genotype at high-throughput are making GWA studies a realistic alternative. As an example of high-throughput, in Phase I of the International HapMap Project over one million SNPs were genotyped in each of 269 DNA samples and in Phase II an additional 4.6 million genotyped SNPs per DNA sample is the goal (International HapMap Consortium, 2005).

Commercially available high-throughput genotype platforms with genome-wide coverage have been made available by companies such as Third Wave, Sequenom, ABI, Illumina, Parallele, Affymetrix, and Perlegen. Each company utilizes a unique genotyping assay. These platforms allow the interrogation of a large number of genetic markers on a sample of subjects with the use of robotic automation, though the cost to do so may be restrictive. An example of such genotyping technologies is Illumina's HumanHap300-Duo and -Duo+ Genotyping BeadChips (www.illumina.com). The HumanHap300-Duo accomodates two DNA samples simultaneously on more than 318,000 tag SNPs selected from Phase I and II of the International HapMap Project. The two sample format ostensibly decreases experimental variability. In regions near a gene or in evolutionarily conserved regions there is an increased density of tag SNPs.

1.3.5 Utility of the International HapMap Project in association studies

When planning a GWA study, determining how many and which markers to genotype is crucial to potential success, which can be aided by utilizing the haplotype map of the International HapMap Project. This haplotype map reveals patterns of LD across the entire human genome (International HapMap Consortium, 2005).

In particular, the aim of the International HapMap Project was to provide tools

for genetic studies (e.g. candidate gene, linkage, and GWA studies) based on the 'indirect' association approach. Collins et al. (1997) proposed this 'indirect' approach, in which a set of genetic markers could be used to test for disease association in genomic regions, and these markers would not necesarily be required to have a functional effect on disease status. Subsequently, causal sequence variants could then be explored in genomic regions where associations with disease have been previously found. On the other hand, the 'direct' approach assesses disease association for each putative causal variant across the entire genome (Risch, 2000). To locate candidate variants would require sequencing the whole genome of many patient samples for a considerable cost (Botstein and Risch, 2003).

The members of the International HapMap Consortium (2003) believe that the indirect approach has the potential of outperforming the direct approach with respect to capturing most human sequence variation, based upon ideas from human population genetics. Kruglyak and Nickerson (2001) claim that about 90% of human sequence variation consists of common variants. Furthermore, these common variants are the result of a single mutation which ocurred at some point in time, so variants in close proximity on the same chromosome are associated with the mutation. The indirect approach capitalizes on these associations by using a small set of variants that represent the LD patterns of common variation in the genome. Thus, it is not necessary to obtain previous knowledge about functional variants in order to scan the regions of interest (International HapMap Consortium, 2003). The hope is that a region or gene associated with disease would be discovered even if the particular genetic marker tested is not the causal variant. Additionally, the amount of genotyping (and hence the overall study cost) would be significantly reduced since a subset of representative common variants rather than the entire set would be genotyped. In terms of GWA studies, Balding (2006) approximates that around 300,000 SNPs would capture a majority of

the common genetic variation in Caucasians and more SNPs for African poplations due to increased genetic diversity.

1.3.6 Qualitative vs. quantitative phenotype

Phenotypes can be measured qualitatively or quantitatively. Qualitative phenotypes are dichotomous, for example, presence or absence of a disease. A quantitative trait is measurable and could contain discrete values (e.g. number of tumors) or could be continuous (e.g. blood pressure). It is thought that the variation in quantitative traits could be explained by genetic and/or environmental factors (Complex Trait Consortium, 2003). Schork and Chakravarti (1996) describe qualitative traits as possibly having multiple genetic and perhaps nongenetic determinants. It could be argued that all traits could be considered quantitative since quantitative variables such as hormones and protein amounts may be involved in disease pathogenesis. On the contrary, alleles and mutations, which may underly disease onset or phenotypic expression, are discrete in nature and thus it is not possible that all traits are quantitative.

The Complex Trait Consortium (2003) state that the same mapping strategies can be used to search for causal loci in monogenic Mendelian disorders as well as for QTLs. They view the classification of genetic effects as a continuum where on one end lies the single gene effect of Mendelian diseases with a dichotomous outcome (i.e. affected or unaffected). On the other end are quantitative traits that are influenced by multiple genes, each with a small effect. In between these two poles are traits that are controlled by multiple loci and possibly environmental determinants resulting in several intermediate phenotypes.

1.3.7 Common vs. rare causal genetic variants

As mentioned in sections 1.2 and 1.3, linkage analysis has been successful in mapping rare genetic variants in monogenic Mendelian disorders whereas association studies are more suited for detecting common alleles in complex diseases. However, two studies have identified common variants using linkage analysis. Human leukocyte antigen was suggested to be involved in type 1 diabetes (Concannon et al., 1998) and apolipoprotein E was shown to play a role in late-onset Alzheimer's disease, though an abundant amount of references such as these may not exist.

A rough guide to variants considered common is having a minor allele frequency above 5% (Balding, 2006). The Common Disease/Common Variant (CD/CV) idea, proposed in the late 1990's (Lander, 1996; Cargill et al., 1999; Chakravarti, 1999), hypothesizes that common genetic variants are responsible for risk of common diseases. Although several reports such as Corder et al. (1993), Bertina et al. (1994), and Altshuler et al. (2000) support the CD/CV hypothesis, the extent to which it holds remains unclear (Balding, 2006). Alternatively, multiple rare genetic variants may contribute to complex diseases.

Rare variants may be more likely discovered in population isolates and founder populations (e.g. Saami of Scandinavia, Laan and Pbo (1997)) due to their extensive LD patterns. For example, it has been shown that there is considerable LD around rare disease mutations in population isolates such as the Finns, Ashkenazi Jews, and Mennonites (Risch, 2000). Moreover, for association tests power improves significantly as LD increases. However, for these populations it may be unlikely to detect common alleles (Wright et al., 1999).

Rare variants may be the result of a recent mutation and therefore restricted to a single founder population or they may be the result of a historical mutation and typically found in one major ethnic group (Risch, 2000). Thus, the same gene in other

populations could be investigated for other functional variants associated with a similar trait. If multiple functional variants are found, then this would strongly suggest a causal relationship between the gene and trait.

On the other hand, common variants are probably present in many different populations consisting of varying ethnicities (Risch, 2000). Positive associations between a candidate variant and trait across multiple populations would imply causality. However, non-reproducibility would not necesarily refute a causal relationship since among populations, gene expression may result in differing degrees of association.

1.3.8 Single SNP tests of association: unrelated cases and controls

When SNPs are genotyped, there are several analytic approaches that can be employed. In the genotype association test of 2 degrees of freedom, cases and controls can be classified in the rows and genotypes in the columns (Balding, 2006). For a diallelic SNP with alleles 'D' and 'd', the possible genotypes are dd (homozygotes), Dd (heterozygotes), and DD (homozygotes). Thus, the resulting contingency table is of dimension 2×3 and contains the respective cross classified counts. To assess the null hypothesis of no association between disease status and genotype, either the Pearson's 2 degree of freedom χ^2 test or Fisher's exact test may be used. The latter is recommended if the contingency table contains small expected cell counts. It is based on enumerating all possible realizations of cell counts given the marginal totals, and so is computationally burdensome. Both tests are readily available in standard statistical software programs.

There is widespread belief (Balding, 2006) that any particular causal SNP will often approximately influence disease risk in an additive fashion. In other words, assuming 'D' is the causal disease allele, the risk of disease for a heterozygote Dd will be intermediate between the homozygous risks of dd and DD, where the risk for homozygotes dd (i.e.

those without any copy of the disease allele) is less than that of homozygotes DD (i.e. those with two copies of the disease allele). The genotype association test described above performs reasonably well in terms of power despite the underlying risks involved. However, the allelic association test of 1 degree of freedom is more powerful than the genotype association test if the genotype risks are additive. The construction of the 2 $\times$ 2 contingency table is as follows. Disease status constitutes the rows and alleles are classified in the columns instead of genotypes, such that each subject contributes two counts to any given cell. In particular, a homozygote dd will be counted twice in the d allele column, similarly for a homozygote DD in the D allele column, and a heterozygote Dd contributes one count to the d column and another to the D column. Pearson's 1 degree of freedom χ^2 statistic or Fisher's exact test may be computed to test the null hypothesis of no association. The main drawback of this approach is that alleles within any given individual must be independent, or in genetics terms, Hardy-Weinberg equilibrium (HWE) must hold in both cases and controls. Due to the assumption of HWE and the observation that risk estimates are not interpretable, Sasieni (1997) does not recommend the 1 degree of freedom allelic association test.

Using the 2 $\times$ 3 contingency table construction as described above, the Cochran-Armitage test (Armitage, 1955) is yet another analytic option. This method tests for a linear trend in the proportion of cases relative to the 'ordered' genotypes dd, Dd, and DD. In the case of the additive genetic disease model, one would expect subjects with two copies of the disease allele (i.e. genotype DD) to exhibit a higher proportion of cases compared to individuals with zero copies (i.e. genotype dd), and those with one copy (i.e. genotype Dd) to have an intermediate proportion of cases. The Cochran-Armitage test is conservative and similar to the genotype association test, does not require the assumption of HWE. It has good power for the additive model, though the farther from the additive model, the more the power diminishes (Balding, 2006).

1.3.9 Haplotype based approaches: motivations and difficulties

Haplotypes have proven to be important in the fine-mapping of Mendelian disorders (Schaid, 2004b). It is now the hope that haplotypes will facilitate the genetic discovery and mapping of common diseases that are polygenic, unlike single-gene Mendelian disorders. Haplotypes have been used in GWA studies, motivated by empirical results suggesting that haplotype 'blocks' define the sequence variation throughout the genome, in which the blocks are more conserved than in other regions (Daly et al., 2001; Jeffreys et al., 2001; Patil et al., 2001; Gabriel et al., 2002).

Furthermore, McVean et al. (2004) offer strong evidence that recombination rates are not uniformly distributed across the genome and that certain regions of the genome are more likely to recombine (i.e. 'hot' spots) whereas other regions are less likely (i.e. 'cold' spots), resulting in areas of weak and strong LD, respectively.

Aside from the observation of haplotype blocks, there are several additional reasons that suggest the utility of haplotypes. There are biological aspects of haplotypes. Previously, genetic markers were widely spaced (Schaid, 2004a), and thus not capturing the DNA sequence regions pertaining to biological function. Presently, the genotyping of SNPs has been at a higher density, such that these genetic markers comprising haplotypes are more representative of regions of biological function. Additionally, in light of the central dogma, DNA sequence variation on a haplotype gives rise to the linear arrangement of amino acids via transcription and translation, which subsequently determines protein folding (Clayton et al., 2004). Furthermore, there are several reports of a 'super-allele' (Schaid, 2004a), which is the result of multiple mutations on the same haplotype that interact with each other to largely influence the observed phenotype (Clark et al., 1998; Drysdale et al., 2000; Hollox et al., 2001).

The use of haplotypes to test for a trait of interest offers some statistical advan-

tages. Schaid (2004a) has reviewed the literature on the statistical power of analyzing haplotypes compared to single markers. It is difficult to directly compare the results from various reports on the statistical power of analyzing haplotypes in contrast to single markers, since investigations differ in their assumptions, though Schaid (2004a) concludes the following. For quantitative traits, if there are more haplotypes than causative SNPs (i.e. SNPs that influence the trait), among all the measured SNPs, than single SNP tests are more powerful than haplotype based tests (Bader, 2001). In addition, Long and Langley (1999) found that despite testing SNPs that were not causal but rather in LD with a QTL, the single SNP tests were still more powerful. On the other hand, for the dichotomous outcome of affecteds and unaffecteds in the case-control setting, haplotype based tests are more powerful when the SNPs are in LD with a causal variant (Akey et al., 2001). As described previously about common complex diseases, several genetic variants may each contribute a portion to disease risk. In terms of detecting these multiple associations, both the single marker and haplotype based approaches lose power, though the latter suffers less (Slager et al., 2000). The haplotype based tests offer the largest benefit in terms of power compared to the single locus tests when the markers are in strong LD with the causal variants and not with each other (Morris and Kaplan, 2002).

For haplotype methods, it is expected that surrounding the causal genetic variant on the haplotypes of affected individuals (i.e. case haplotypes), there are significantly longer stretches of DNA identical by descent (IBD) compared to randomly selected haplotypes (Nolte, 2002). This is due to the inreased relatedness of the regions around the predisposing mutation in the case haplotypes in contrast to randomly selected haplotypes. Furthermore, the haplotype is the result of genetic drift and past mutational and recombinational events, i.e. it is a reflection of evolution as it is transmitted throughout generations. Therefore, approaches that measure the amount of sharing among

haplotypes account for the evolutionary history of the sample (Beckmann et al., 2005), whereas single locus tests do not.

Despite these reasons to employ haplotypes in gene mapping strategies, there are several difficulties. First, including rare haplotypes in analyses inevitably increases the degrees of freedom and thus reduces power (Balding, 2006). Second, in most cases haplotypes are not directly measured and must be statistically inferred using unphased genotype data. However, in order to empirically determine haplotype phase in the laboratory, molecular haplotyping methods have been developed such as pyrosequencing (Odeberg et al., 2002), intracellular ligation (McDonald et al., 2002), and clone-based systematic haplotyping (Burgtorf et al., 2003), to name a few. Haplotyping methods are not widely used because a relatively large number of samples cannot be processed in a relatively short period of time, they are costly to implement, and technical issues have not been fully addressed (Niu, 2004). Therefore, the current viable alternative is to infer haplotypes using algorithms, though assessing the overall statistical significance is problematic when there are phase uncertainties (Balding, 2006).

Third, the standard use of automated genotyping procedures unavoidably results in ambiguities when scoring genotypes (Kang et al., 2004), which could lead to genotyping errors since almost all genotyping machines assign a genotype despite the presence of ambiguities (Niu, 2004). Thus, because of genotyping error, two haplotypes may be similar yet not completely identical though they may share a common ancestor. The dissimilarity may also be due to recombination events or mutations (Lange and Boehnke, 2004). If both contain the causal variant, their separate effects on disease risk may not be detectable unless they are jointly examined.

1.4 Definitions

Alleles: "alternate forms of a gene or genetic locus that differ in DNA sequence" (Ellsworth and Manolio, 1999a)

Base pair: The pair of nitrogenous bases, consisting of a purine linked by hydrogen bonds to a pyrimidine, that connects the complementary strands of DNA or of hybrid molecules joining DNA and RNA. The base pairs are adenine-thymine and guanine-cytosine in DNA, and adenine-uracil and guanine-cytosine in RNA.

CentiMorgan (cM): "unit for expressing relative distance between genes or markers on a chromosome equal to 1% recombination; one cM corresponds roughly to a physical distance of one megabase (Mb, one million base pairs)" (Ellsworth and Manolio, 1999b)

Chromosome: A threadlike linear strand of DNA and associated proteins in the nucleus of eukaryotic cells that carries the genes and functions in the transmission of hereditary information.

Complex trait: "...refers to any phenotype that does not exhibit classic Mendelian recessive or dominant inheritance attributable to a single gene locus." (Lander and Schork, 1994)

Crossover: see recombination

Deoxyribonucleic acid (DNA): "...a macromolecule that carries genetic information and represents the molecular basis of heredity...There are four common nitrogenous bases in DNA: two purines—adenine (A) and guanine (G) and two pyrimidines—cytosine (C) and thymine (T). The double-stranded molecule is twisted in the form of a helix with a constant width maintained by restrictions to base pairing such that A only pairs with T and G only pairs with C." (Ellsworth and Manolio, 1999a)

Founder populations: "Populations that have been derived from a limited pool of individuals within the last 100 or fewer generations." (Hirschhorn and Daly, 2005) "...and have undergone a demographic expansion with negligible migration afte foun-

dation." (Bourgain et al., 2000)

Gene: A hereditary unit consisting of a sequence of DNA that occupies a specific location on a chromosome and determines a particular characteristic in an organism. Genes undergo mutation when their DNA sequence changes.

Genetic drift: "The random fluctuation in allele frequencies as genes are transmitted from one generation to the next." (Cardon and Bell, 2001)

Genotype: The combination of alleles located on homologous chromosomes that determines a specific characteristic or trait.

Genotyping: The process of determining the genotype of an individual by the use of biological assays.

Haplotype: "The specific set of alleles observed on a single chromosome, or part of a chromosome..." (International HapMap Consortium, 2003)

Haplotyping: The process of determining the haplotypes of an individual by the use of biological assays.

Hardy-Weinberg equilibrium: "Holds at a locus in a population when the two alleles within an individual are not statistically associated." (Balding, 2006) or "The binomial distribution of genotypes in a population, such that frequencies of genotypes AA, Aa and aa will be p^2, 2pq, and q^2, respectively, where p is the frequency of allele A, and q is the frequency of allele a. Hardy-Weinberg equilibrium applies in a population when there are no factors such as migration or admixture that cause deviations from p^2, 2pq, and q^2." (Hirschhorn and Daly, 2005)

Heterozygote: An organism that has different alleles at a particular gene locus on homologous chromosomes.

Homozygote: An organism that has the same alleles at a particular gene locus on homologous chromosomes.

Identity by descent (IBD, see identity by state): "...the identity of two stretches

of DNA due to inheritance from a common ancestor without recombinations and mutations ..." (Nolte and te Meerman (2002)) or "Alleles that trace back to a shared ancestor. For sibs, refers to inheritance of the same allele from a given parent." (Risch, 2000)

Identity by state (IBS, see identity by descent): "Alleles are IBS if they are simply of the same type." (Schork and Chakravarti, 1996)

Linkage: "the proximity of multiple genes or markers on the same chromosome reduces the probability that recombination events will occur between them and increases the probability that certain combinations of alleles at these genes or markers will be inherited together as a linkage group or haplotype" (Ellsworth and Manolio, 1999b)

Linkage disequilibrium (see linkage equilibrium): "the nonrandom transmission from parents to offspring of alleles from genes or markers that are located on the same chromosome. Because alleles at tightly linked loci are often inherited together, linkage disequilibrium is useful for detecting regions of the genome that historically have been inherited as a linkage group and may help identify the approximate location of genes that contribute to disease (*fine mapping*)" (Ellsworth and Manolio, 1999b)

Linkage equilibrium (see linkage disequilibrium): the converse of linkage disequilibrium

Locus (plural: loci): "a position on a chromosome or segment of DNA, usually used in reference to a gene or genetic marker" (Ellsworth and Manolio, 1999b)

Marker: "an identifiable physical location on a chromosome or DNA segment useful in genome mapping and linkage analysis. Numerous types of sequences are considered markers including functional genes, portions of expressed sequences (expressed sequence tags or ESTs), short DNA segments that are detected by PCR (sequence-tagged sites or STSs), microsatellites, restriction fragment length polymorphisms (RFLPs), and single-nucleotide polymorphisms (SNPs)" (Ellsworth and Manolio, 1999b)

Megabase (Mb): "unit of physical measurement for nucleic acids equal to one million base pairs, roughly equivalent to a genetic distance of one centiMorgan (cM)" (Ellsworth and Manolio, 1999b)

Mendelian inheritance: A set of primary tenets relating to the transmission of hereditary characteristics from parent organisms to their children; it underlies much of genetics.

Minor allele: the less abundant allele

Minor allele frequency: the frequency of the less abundant allele

Mutations: "occasional errors that occur during DNA replication" (Ellsworth and Manolio, 1999a)

Penetrance: The proportion of individuals carrying a particular variation of a gene (allele or genotype) that also express an associated trait (phenotype).

Phenotype (see genotype): any observable characteristic or trait of an organism.

Polymorphism: "the existence of multiple forms of a gene or genetic locus (alleles) that differ in DNA sequence" (Ellsworth and Manolio, 1999a)

Population admixture: see population stratification

Population stratification (see population admixture and structure): "The presence of multiple subgroups with different allele frequencies within a population. The different underlying allele frequencies in sampled subgroups might be independent of the disease within each group, and they can lead to erroneous conclusions of linkage disequilibrium or disease relevance." (Cardon and Bell, 2001)

Population structure: see population stratification

Quantitative trait locus (QTL): "a genetic factor believed to influence a quantitative trait such as blood pressure lipoprotein levels" (Ellsworth and Manolio, 1999b)

Recombination: "process by which homologous chromosomes physically exchange segments of DNA (also known as crossing-over)" (Ellsworth and Manolio, 1999c)

Single nucleotide polymorphism (SNP, pronounced 'SNiP' or 'S' 'N' 'P'): "polymorphism where a single base substitution has created two forms of a DNA sequence that differ by a single nucleotide - currently of great interest for locating genes associated with complex diseases" (Ellsworth and Manolio, 1999c)

tag SNP: "Single nucleotide polymorphisms that are correlated with, and therefore can serve as a proxy for, much of the known remaining common variation in a region." (Hirschhorn and Daly, 2005)

Transmission disequilibrium test (TDT): "A family-based test for association that is immune to population stratification. The transmission of alleles from heterozygous parents to affected offspring is compared to the expected 1:1 ratio." (Hirschhorn and Daly, 2005)

CHAPTER 2

USING PUBLIC CONTROL GENOTYPE DATA TO INCREASE POWER AND DECREASE COST OF CASE-CONTROL GENETIC ASSOCIATION STUDIES

2.1 Introduction

Large-scale commercial genotyping platforms have facilitated the identification of numerous common single nucleotide polymorphisms (SNPs) that are associated with complex genetic diseases. The newest commercial genotyping platforms now contain over 1 million SNPs spread across the human genome. While the cost per genotype on these platforms have decreased considerably over the past several years, the cost per sample

remains prohibitive for many scientific investigators who are interested in performing a genome-wide association (GWA) study using their own samples.

The high-cost of GWA studies has lead to the utilization of multi-stage study designs, a strategy routinely used in clinical trials. Two-stage genotyping designs typically involve genotyping a fraction of the entire sample on a commercial genotyping platform containing all SNPs of interest in stage 1, performing systematic tests of association using stage 1 samples, and genotyping stage 2 samples on only the SNPs of greatest interest as determined in stage 1 (Satagopan et al., 2002). Two-stage genotyping designs have been shown to maintain power comparable to a single-stage study employing all samples while substantially decreasing overall genotyping costs (Satagopan et al., 2002; Satagopan et al., 2004; Skol et al., 2006; Thomas et al., 2004). The data collected from the second stage of a two-stage GWA study is either analyzed separately as a replication-based sample or the data is combined with data from the first stage and the combined data is analyzed jointly. The replication-based approach requires a less stringent significance threshold, due to a smaller multiple test correction factor that is based on only the number of markers followed up in stage 2 samples, than the joint analysis approach, that uses a correction factor that accounts for the entire number of markers studied in stage 1. The joint analysis approach benefits from using all of the available data as opposed to just the data from samples genotyped in the second stage.

A recent alternative approach for reducing the cost of a large-scale case-control genetic association study is to use freely available genotype data from previous genome-wide association scans as control data in the current study. The effective use of a common control dataset for comparison with multiple case datasets for different phenotypes was illustrated by the Wellcome Trust's Case Control Collaboration (WTCCC) GWA study on 14,000 cases of seven common diseases and 3,000 shared controls (Wellcome Trust Case Control Consortium, 2007). In this study, based on British subjects of

European descent, the WTCCC identified 24 independent associations ($p < 5$ x 10-7) for bipolar disorder, coronary artery disease, Crohns disease, rheumatoid arthritis, type 1 diabetes and type 2 diabetes using 2,000 independent cases for each disorder. The WTCCC demonstrated that utilizing a common control dataset can be a powerful and cost effective approach for performing future GWA studies.

For investigators that have collected a well-matched group of cases and controls who wish to preserve many of the benefits of their sample collection design, we describe a two-stage replication-based case-control genetic association study design that uses free genotype data from public controls in stage 1, well-matched study controls in stage 2, and study cases distributed over stages 1 and 2. We compare the power and relative cost of our two-stage approach to single-stage approaches that strictly use either free public control genotype data or genotype data from study controls and to the single-stage approach that combines public and study controls. We discuss the advantages and limitations of each of the four sampling designs and show that the proposed two-stage replication-based study design using both public and study controls is robust to high proportions of mismatched public controls and batch genotype effects that can result from genotyping samples different populations at different times.

2.2 Methods

We assumed an investigator had a sample of N_A study cases, N_U study controls and access to free genotype data on N_{PU} public controls. We further assumed that study controls may or may not be screened for disease and that public controls had not been screened for disease. We performed a series of calculations over a range of alternative models comparing the power achieved in an association study using four different sampling approaches: 1) a single-staged association study that used all N_A study cases

and N_U study controls; 2) a single-staged association study that utilized all N_A study cases and N_{PU} public controls; 3) a two-staged replication-based study that used all N_{PU} public controls in stage 1, all N_U study controls in stage 2 and all N_A cases apportioned between stages 1 and 2; 4) a single-staged association study that used all N_A cases and combined all N_U study and N_{PU} public controls. We assumed an underlying multiplicative genetic mode-of-inheritance risk model for a bi-allelic locus with alleles D and d and corresponding allele frequencies of f_D and f_d, respectively. For each alternative model, we set the population frequency of the susceptibility allele D in the general population, the prevalence (K) of the disease in the population, and the locus specific genetic relative risk (GRR) = Pen(DD) / Pen(Dd) = Pen(Dd) / Pen(dd), where Pen(dd), Pen(Dd), and Pen(DD) were the penetrances for the dd, Dd, and DD genotypes, respectively. Consistent with many genetic power calculators, our power calculations are for the main effects of a directly genotyped locus and, as such, do not rely on additional assumptions regarding the extent of linkage disequilibrium between this locus and an untyped causal locus. All power analyses were programmed into the freely available statistical software R version 2.4.1 (R Development Core Team, 2006).

2.2.1 Single-stage Power Calculations

Assuming Hardy-Weinberg equilibrium in the general population from which the cases and controls were selected, we used our model assumptions (allele frequencies, disease prevalence and GRR) to calculate the penetrance functions and we used Bayes' theorem to ascertain the conditional probability of each genotype given affection status, P_{ji}, where j = 0 (cases), 1 (controls) and i = 0 (dd), 1 (Dd), 2 (DD). Namely, for the cases these probabilities were P_{00} = Pr(dd | case), P_{01} = Pr(Dd | case), and P_{02} = Pr(DD | case) and for the unaffected (screened) controls the probabilities were P_{10} = Pr(dd | unaffected control), P_{11} = Pr(Dd | unaffected control), P_{12} = Pr(DD | unaf-

fected control). We assume no disease misclassification among study cases or screened study controls. Derivations of the conditional genotype probabilities are provided for the multiplicative model in the Supplementary materials. For unscreened and public controls, the genotype probabilities for controls were set to the genotype probabilities in the general population, namely $P_{10} = f_d^2$, $P_{11} = 2f_d f_D$, $P_{12} = f_D^2$, since affection status was not assumed to be known.

We calculated asymptotic power for the Cochran-Armitage trend test (Armitage, 1955; Cochran, 1954) by specifying the non-centrality parameter based on work by Chapman and Nam (1968) and we set the vector of scores to x = (0, 1, 2) for genotypes (dd, Dd, DD), respectively (Slager and Schaid, 2001). In particular, the non-centrality parameter, explicitly stated by Ahn et al. (2007), was

$$\lambda = N_A N_U \frac{\left[\sum_{i=0}^{2} x_i \left(P_{0i} - P_{1i}\right)\right]^2}{\sum_{i=0}^{2} x_i^2 \left(N_A P_{0i} + N_U P_{1i}\right) - \left[\sum_{i=0}^{2} x_i \left(N_A P_{0i} + N_U P_{1i}\right)\right]^2 \Big/ \left(N_A + N_U\right)} \tag{2.1}$$

where N_A and N_U (or optionally N_{PU}) were the sample sizes of the cases and screened (or public) controls, respectively, x_i was the score for the i-th genotype (i = 0, 1, 2 for genotypes dd, Dd, DD), and P_{0i} and P_{1i} were the probabilities of the i-th genotype for the cases and controls, respectively. Power was then taken to be $1 - \beta$, where was the type II error and was the cumulative distribution function of the non-central χ^2 distribution with 1 degree of freedom and non-centrality parameter λ, evaluated at the $100(1 - \alpha_{\text{Bonferroni}})$ percentile of the central χ^2 distribution with 1 degree of freedom. For single-stage designs, the overall family-wise error rate was set to $\alpha = 0.05$ by using a Bonferroni corrected significance threshold $\alpha_{\text{Bonferroni}} = 0.05/\text{M}$, where M is the number

of markers evaluated.

2.2.2 Two-stage Power Calculations

Using the formulas described above for one-stage power, we calculated power for a replication-based two-stage design. For a two-stage replication-based design, the overall power for a SNP was simply calculated as the product of the power for the first stage times the power of the second stage. Following the notation in Skol et al. (2006), the power for the first-stage was calculated using a significance threshold defined as the proportion of markers followed in stage 2, π_{markers}. Power for the second-stage was calculated using a significance threshold (assuming a two-sided test) equal to $\alpha/(M \cdot \pi_{\text{markers}})$, i.e. the Bonferroni corrected cutoff, where M was the number of markers typed and interrogated in stage 1. Setting the significance cutoff at markers in stage 1 on average resulted in markers being the type I error. Similar to Skol et al. we also calculated the power for a one-sided test in stage 2 samples, requiring the effect for the SNP to be in the same direction in both stage 1 and stage 2 samples.

While Skol et al. allowed markers to be any possible value, we restricted the number of SNPs for follow-up analysis in stage 2 to be values that approximate numbers that would typically be considered given todays currently available commercial genotyping platforms. Namely, we considered follow-up platforms of size 100, 375, 1,500, 7,500, and 16,500 SNPs. For each follow-up genotyping platform, we then found the optimal proportion of cases, cases, to be genotyped in stage 1 that optimized the power of the two-stage design. Specifically, we used the "optimize" function in R to search for the maximum power in the continuous space of cases. This method combines the golden section search and successive parabolic interpolation algorithms.

2.2.3 Single-stage power calculation for combined public and screened study controls

We used simulations to estimate the power of the single-stage study design that compared allele frequencies between study cases and the combined sample of public and screened study controls. Specifically, we simulated 10,000 data sets for each model condition and used the Cochran-Armitage trend test, implemented in R, to test for association between marker and disease. Similar to the other single-stage designs, the overall family-wise error rate was set to $\alpha = 0.05$ by using a Bonferroni corrected significance threshold $\alpha_{\text{Bonferroni}} = 0.05/\text{M}$, where M was the number of markers evaluated.

2.2.4 Examples of Power Approximations for 1- and 2-Stage Designs

We calculated power for three models to demonstrate the difference in power between the competing approaches. For all three models, we assumed a multiplicative model with a GRR = 1.3, and a susceptibility allele frequency $f_D = 0.3$ in the general population. In addition, for all three models we performed the calculations assuming study controls (in stage 2) have or have not been screened for disease. Model 1 was a GWA scan on M = 500,000 SNPs for a study sample of N_A = 2,000 study cases and N_U = 2,000 study controls. Model 2 was identical to Model 1, except that there were fewer study controls, N_U = 1,000. Model 3 was designed to mimic a targeted follow-up study to a previous GWA study. For Model 3, M = 7,500 and $N_A = N_U$ = 1,250. For all three models we considered a wide range of disease prevalence values of K = 1 x 10^{-4}, 0.01, 0.05, 0.1, 0.25, and 0.5 and we assumed available genotype data on samples of N_{PU} = 1,000, 3,000, 5,000 and 10,000 public controls. We calculated power for the single-stage designs using only study controls, only public controls, or both control samples

combined. We also calculated the power for the optimal two-stage replication designs using one- and two-sided hypothesis tests in stage 2. For each optimal two-stage model we define the optimal platform and proportion of cases, cases, genotyped in stage 1. Finally, in order to test how power the 1- and 2-stage designs are impacted by different possible combinations of disease allele frequency, disease prevalence, and GRR, we calculated power for Model 1 (assuming N_{PU} = 5,000) using disease susceptibility allele frequencies of f_D = 0.1 and 0.5, disease prevalences of K = 0.01, 0.1 and 0.25, and GRRs ranging from 1.1 to 1.5.

In the above power calculations, for the two-staged replication approach we chose the follow-up platform and proportion of cases genotyped in stage 1 that optimized power under a specific alternative hypothesis, namely, the relative risk and disease allele frequency (in the general population) were explicitly defined. In practice the true alternative model is unknown. A desirable quality of any two-stage approach is that the optimal choice of follow-up platform and the optimal proportion of cases genotyped on the follow-up platform are robust to the underlying relative risk and disease allele frequency. We performed additional power calculations to assess the robustness of the choice of follow-up platform and the proportion of cases, cases, genotyped on the follow-up platform across a range of alternative models. Specifically, assuming a GWA study on M = 500,000 SNPs using N_A = 2,000 study cases, N_U = 2,000 screened study controls and N_{PU} = 5,000 public controls for a multiplicative trait with a prevalence K = 0.1, we calculated the maximum power and corresponding proportion of cases genotyped in stage 1, across a range of relative risks (GRR = 1.25-1.5) and disease allele frequencies (f_D = 0.1, 0.3, and 0.5) based on follow-up platforms containing 100, 375, 1,500, 7,500 and 16,500 SNPs. In addition, assuming a relative risk of 1.3 and disease allele frequency of 0.3, we calculated power across a range of proportion of cases, cases, genotyped for each of the 100, 1,500, 7,500, and 16,500 SNP follow-up

platforms to assess the decrease in power when using a higher or lower proportion of cases in stage 1 compared to the optimal proportion for each platform.

In the supplementary material, we performed additional power calculations using the general model (co-dominant) test of association (two-degree-of-freedom Chi-square test) under the same multiplicative alternative hypothesis models we considered for the Cochran-Armitage trend test. In addition, power was also calculated for several dominant and recessive inheritance models using the single-degree-of-freedom Chi-square test.

2.2.5 Impact on Power of Ancestrally Poorly-Matched Public Controls and Batch Genotype Effects

In the previous calculations, we did not consider the impact of ancestrally poorly-matched public controls and batch genotype effects on power that can occur when genotyping samples of cases and public controls from different populations at different times. We evaluated the impact of these factors for a study design that included 2,000 study cases, 2,000 study controls, and 5,000 public controls for a multiplicative disease model with susceptibility allele frequency = 0.3, K = 0.10 and GRR = 1.3. For ancestrally poorly-matched public controls (with respect to our study cases), we measured the reduction in power by decreasing the effective sample size of the public control sample. Specifically, for the purpose of these calculations, we have assumed that a fraction (we considered a range from 0% to 90%) of public controls will be removed from consideration after genotyping study cases (when comparisons of ancestry can be made between study cases and public controls using genome-wide data) and prior to performing association testing. We have additionally assumed that the proportion of cases genotyped in stage 1 of our two-stage replication design is optimized and chosen prior to the removal of any public controls. Power calculations were also performed

for the two one-stage designs that utilize public controls after eliminating ancestrally poorly-matched public controls.

To help assess the impact of batch genotype effects on our proposed two-stage design we calculated power using more stringent significance thresholds for stage 1. We assumed that batch genotype effects in stage 1 would lead to an excess of SNPs, under the null hypothesis, with low p-values and that the SNP associated with disease was not subject to batch genotype effects. The impact of batch genotype effects under these assumptions was that truly associated SNPs were required to reach a higher significance level in stage 1 than anticipated in order to be included in stage 2 genotyping. We calculated power in stage 1 of our two-stage replication design by varying the magnitude of the departure of the required significance threshold from markers in stage 1 (p-value required for a SNP to be genotyped in stage 2) to be between 0.99 x π_{markers} and 0.1 x π_{markers}. The proportion of cases genotyped in stage 1 was optimized under the erroneous assumption of no batch genotype effects (i.e. markers was assumed to be the significance threshold required for a SNP to be subsequently genotyped in stage 2). Power calculations that included batch genotype effects were not performed for the three one-stage designs.

2.2.6 Example of Genotyping Costs for Different Genotype Sampling Strategies

To understand the financial impact of the different genotyping sampling strategies, we estimated the relative cost of each genotype sampling design for a GWA study based on M = 500,000 SNPs using N_A = 2,000 study cases, N_U = 2,000 screened study controls and N_{PU} = 5,000 public controls. We assumed a multiplicative trait with a prevalence K = 0.1, GRR = 1.3 and f_D = 0.3 (Model 1). We calculated the relative costs of performing the three single-stage studies that used either study or public controls or

both. For these single-stage sampling designs, all samples were assumed to be genotyped on all 500,000 SNPs; genotype data for public controls were assumed to be available at no expense. In addition, we calculated the relative cost of the optimal (highest power) replication-based two-stage study design for each follow-up platform. For the purpose of our calculations, we assumed the Illumina Human660W-Quad platform would be used for genotyping 500,000 viable SNPs in stage 1 and Illuminas GoldenGate 96, 384 and 1,536 SNP panels and Illuminas Custom iSelect Infinium 7,600 and 16,720 SNP panels would be used as the follow-up platforms for stage 2. Given that genotyping costs are constantly changing, rather than use dollar amounts, we report the relative cost of genotyping based on the most current prices. Using the cost of genotyping 500,000 SNPs in a GWAS as a baseline, the relative cost of genotyping 16,000, 7,500, 1,500, 375 and 100 SNPs were assumed to be 1/2, 1/3, 1/5, 1/10 and 1/12 of the cost, respectively, based on the most recent genotype prices at the CIDR genotyping facility (www.cidr.jhmi.edu/pricing.pdf).

Skol et al. (2006) demonstrated that a joint analysis two-stage study design could effectively achieve equivalent power to a single-stage study for a fraction of the cost. Consequently, for the three single-stage sampling designs, we also estimated the relative cost of performing a joint analysis two-stage association study for each follow-up platform. For each combination of sampling design and follow-up platform, we identified the least expensive joint analysis two-stage sampling design that obtained an estimated power within 0.01 of the power obtained from the corresponding single-stage study. For the sampling design that used only public controls, cases were to be divided and genotyped in stages 1 and 2 while all public controls were assumed to be available in stage 1. For the sampling design that included both study and public controls, all study controls were assumed to be genotyped in stage 2, and all public controls were assumed to be available in stage 1. Cases were divided and genotyped in stages 1 and

2. For each study design, we simulated 50,000 replicate data sets to determine the optimal partitioning of study samples into stages 1 and 2 that resulted in the lowest total cost while preserving statistical power.

2.3 Results

We performed power calculations for a range of study designs and disease models. Power is described for the frequency of the risk allele in the general population (the frequency of the risk allele in cases and study controls for different values of K are provided in the table footnotes). Our results showed that utilizing free genotype data from public controls increases statistical power when the number of available public controls is sufficiently large over studies that do not include these data. As expected, combining screened study controls with public control genotype data increased power over sampling designs that included just one or the other for all models considered, regardless of the underlying disease prevalence (Table 2.1). The single-stage study design based solely on public controls had greater power than the single-stage study design based solely on screened study controls for many alternative models when the number of public controls was greater than the number of study controls. However, when the population prevalence of disease was high ($K > 0.25$), the single-stage study design using screened study controls had, in some instances, greater power than the single-stage study using public controls, even when the number of public controls was large. Overall, the same general patterns of results were observed when varying GRR and frequency of the disease susceptibility allele (Supplementary Figure 2.2), when analyzing the genotype data using a general (co-dominant) 2-df inheritance model (Supplementary Table 2.7), and when considering dominant or recessive genetic inheritance models (Supplementary Tables 2.8 and 2.9, respectively).

Power for the proposed replication-based two-stage design was typically greater than the power of the one-stage design based only on study controls for most genetic models provided the prevalence of the disease was not high ($K < 0.25$). For example, for Model 1, assuming screened (unaffected) study controls and a disease prevalence of $K = 0.05$, we had power equal to 0.68 when using only screened study controls. Power increased to 0.69, 0.81, 0.86 and 0.88 when applying our proposed two-staged replication-based approach (with a one-sided test in stage 2) when including available genotype data from 1,000, 3,000, 5,000 and 10,000 public controls, respectively, in stage 1 (Table 2.1). As seen in Model 2, the difference in power between the two-stage replication approach and the single-stage study that used only study controls was more dramatic when the number of available study controls was only half as large as the number of cases. Gains in power were also observed when the initial platform in stage 1 contained only 7,500 SNPs (Model 3), as might be used in a more focused follow-up study of previous GWA scans. Compared to studies using screened study controls, power noticeably decreased for studies using unscreened study controls when $K > 0.05$ (Supplementary Table 2.6). However, the loss in power for the two-stage replication-based approach using unscreened study controls was less dramatic than the drop experienced by the single-stage study based solely on unscreened study controls.

Our results showed that the optimal choice of follow-up platform and proportion of cases used in stage 1 for our proposed two-stage replication sampling design are robust across a range of different possible alternative models. In Table 2.2, we observed that the smallest follow-up platform, containing 100 SNPs, consistently provided the greatest power compared to the other follow-up platforms over the considered range of GRRs and disease allele frequencies, though the differences in maximum power between the different follow-up platform choices was, in most cases, modest. We also noted that the optimal choice of the proportion of cases, cases, to be genotyped in stage 1 varied

considerably between the different platforms (as expected, a larger proportion of cases were necessary to be genotyped in stage 1 for the smallest follow-up platform) but, importantly, varied little within a given platform across the considered range of GRRs and disease allele frequencies. In fact, we noted that for a given follow-up platform, the optimal choice of cases was also robust to analytic strategy (i.e. similar optimal values of cases were observed for the general 2-df test as for the trend test) (Supplementary Table 2.10) and genetic inheritance model (i.e., similar optimal values of cases were also observed for the dominant and recessive models) (Supplementary Tables 2.11 and 2.12). In Figure 2.1, we observed that for each platform, the power dropped very modestly when the choice of the proportion of cases to be genotyped in stage 1 was within 0.05 of the optimal choice. Together these results suggest that it is reasonable to choose an optimal two-stage replication-based study design, namely the choice of follow-up platform and the proportion of cases, cases, to be genotyped in stage 1, based on a specific genetic models and that power should be robust to this choice across a range of alternative genetic models.

We have also demonstrated that our two-stage replication-based study design using public controls is robust to high proportions of ancestrally mismatched public controls (that would have to be eliminated prior to data analyses) and batch genotype effects in stage 1. Specifically, even when eliminating 50% of public controls due to poorly-matched ancestry with study cases, power of the two-stage design was greater than that for the single-stage design based solely on study controls for the model we considered across all stage 2 follow-up platforms (Table 2.3). Interestingly, the larger follow-up platforms were noticeably more robust than the smaller follow-up platforms with respect to the removal of mismatched controls. The single-stage study design that includes only public controls was most strongly impacted by removal of public controls due to ancestry mismatching while the single-stage study design that includes both public and study

controls maintained the greatest power versus all other study designs. Our two-stage replication-based study design was also robust to the increased significance threshold in stage 1 due to batch genotype effects (Table 2.4). Of note, increasing the stage 1 significance threshold required for a SNP to be genotyped in stage 2 by a factor of 2 (0.50 x π_{markers}) had only a small impact on power. Power remained relatively strong even when requiring an order of magnitude higher level of statistical significance in stage 1 (0.10 x π_{markers}) for a SNP to be subsequently genotyped in stage 2.

In addition to increased power, in Table 2.5 we illustrate that substantial cost savings can be achieved for a GWA study when including public controls. We compared the relative cost of one- and two-stage study designs that include study controls, public controls or both. As expected, the most expensive study designs were the one-stage study designs that genotyped all samples (excluding public controls – which provide genotype data at no expense) on all SNPs. Significant cost savings were observed when using the joint-analysis-based two-stage design described by Skol et al. (2006). For example, when utilizing the joint-analysis-based two-stage design following-up the top 1,500 SNPs (corresponding to the 1,536 SNP Illumina GoldenGate custom panel) in stage 2, a 36%, 44% and 60% cost savings was achieved relative to the corresponding one-stage design for sample designs that included only study controls, only public controls and both study and public controls, respectively. The total cost of our proposed replication-based two-stage design was consistently less than the joint-analysis two-stage designs for sample designs that included only study controls or both public and study controls. The study design that included only public controls in a two-stage joint analysis was the least expensive. In addition to having the lowest power, the sampling design that included only study controls was substantially more expensive than any other sampling design.

TABLE 2.1: Power of the Cochran-Armitage trend test for 1- and 2-stage study designs across a range of sample sizes, SNPs in stage 1, and disease prevalences. Study controls are assumed to be screened and disease free.

		Number of Public Controls								
		2,000			5,000			10,000		
K[a]	Study Controls Only	Public Controls Only	2-Stage[b]	Study + Public Controls	Public Controls Only	2-Stage[b]	Study + Public Controls	Public Controls Only	2-Stage[b]	Study + Public Controls
2,000 Cases and 2,000 Screened Study Controls / 500,000 SNPs										
0.0001	0.57	0.56	0.71	0.85	0.90	0.82	0.94	0.97	0.84	0.97
0.01	0.59	0.56	0.72	0.86	0.90	0.82	0.94	0.97	0.85	0.97
0.05	0.68	0.56	0.76	0.89	0.90	0.86	0.96	0.97	0.88	0.98
0.1	0.79	0.56	0.82	0.92	0.90	0.89	0.96	0.97	0.91	0.98
0.25	0.98	0.56	0.95	0.99	0.90	0.97	0.99	0.97	0.97	0.99
0.5	1.00	0.56	1.00	1.00	0.90	1.00	1.00	0.97	1.00	1.00
2,000 Cases and 1,000 Screened Study Controls / 500,000 SNPs										
0.0001	0.19	0.56	0.57	0.76	0.90	0.66	0.93	0.97	0.69	0.97
0.01	0.20	0.56	0.58	0.77	0.90	0.68	0.92	0.97	0.70	0.97
0.05	0.26	0.56	0.63	0.80	0.90	0.72	0.93	0.97	0.75	0.97
0.1	0.36	0.56	0.68	0.83	0.90	0.78	0.94	0.97	0.81	0.98
0.25	0.74	0.56	0.84	0.92	0.90	0.91	0.97	0.97	0.93	0.98
0.5	1.00	0.56	0.99	1.00	0.90	1.00	1.00	0.97	1.00	1.00
1,250 Cases and 1,250 Screened Study Controls / 7,500 SNPs										
0.0001	0.44	0.63	0.59	0.78	0.86	0.62	0.88	0.92	0.64	0.91
0.01	0.45	0.63	0.60	0.78	0.86	0.64	0.87	0.92	0.65	0.91
0.05	0.53	0.63	0.65	0.80	0.86	0.68	0.88	0.92	0.70	0.92
0.1	0.63	0.63	0.70	0.84	0.86	0.74	0.90	0.92	0.76	0.93
0.25	0.91	0.63	0.93	0.93	0.86	0.93	0.94	0.92	0.93	0.95
0.5	1.00	0.63	1.00	1.00	0.86	1.00	0.99	0.92	1.00	0.98

[a] Population prevalence of disease

[b] Optimal 2-stage replication design using all public controls in stage 1 and all screened controls in stage 2, 2-sided test in stage 1 and 1-sided test in stage 2

Risk allele frequency in general population (f_D) = 0.3, genetic relative risk (GRR) = 1.3 assuming a multiplicative model, overall type I error (α) = 0.05

f_D = 0.3 corresponds to f_D = 0.358 in cases for all K and f_D = 0.300, 0.299, 0.297, 0.293, 0.278, 0.228 in screened controls for K = 0.0001, 0.01, 0.05, 0.1, 0.25, and 0.5, respectively

TABLE 2.2: Power for the Cochran-Armitage trend test and the proportion of cases in stage 1 that optimizes power (in parenthesis) in a two-stage replication-based GWA study with 2,000 Cases / 5,000 public controls (stage 1) / 2,000 screened controls (stage 2). Power calculated for one-sided hypothesis test in stage 2.

	Genetic Relative Risk						
f_D[a]	1.20	1.25	1.30	1.35	1.40	1.45	1.50
Follow-up Platform: 16,500							
0.1	0.01	0.05	0.17	0.39	0.63	0.82	0.93
	(0.32)	(0.30)	(0.29)	(0.28)	(0.27)	(0.27)	(0.27)
0.3	0.19	0.53	0.84	0.97	1.00	1.00	1.00
	(0.29)	(0.27)	(0.27)	(0.28)	(0.29)	(0.30)	(0.31)
0.5	0.27	0.64	0.90	0.98	1.00	1.00	1.00
	(0.28)	(0.27)	(0.28)	(0.29)	(0.30)	(0.31)	(0.32)
Follow-up Platform: 7,500							
0.1	0.01	0.06	0.19	0.41	0.65	0.84	0.94
	(0.36)	(0.35)	(0.33)	(0.32)	(0.31)	(0.31)	(0.31)
0.3	0.20	0.55	0.85	0.97	1.00	1.00	1.00
	(0.33)	(0.32)	(0.31)	(0.32)	(0.33)	(0.33)	(0.34)
0.5	0.28	0.66	0.91	0.98	1.00	1.00	1.00
	(0.33)	(0.32)	(0.32)	(0.33)	(0.33)	(0.34)	(0.35)
Follow-up Platform: 1,500							
0.1	0.01	0.07	0.21	0.45	0.69	0.86	0.95
	(0.45)	(0.44)	(0.43)	(0.41)	(0.41)	(0.40)	(0.4)
0.3	0.22	0.59	0.87	0.97	1.00	1.00	1.00
	(0.42)	(0.41)	(0.40)	(0.40)	(0.41)	(0.41)	(0.41)
0.5	0.31	0.69	0.92	0.99	1.00	1.00	1.00
	(0.42)	(0.41)	(0.41)	(0.41)	(0.41)	(0.42)	(0.42)
Follow-up Platform: 100							
0.1	0.02	0.08	0.24	0.49	0.73	0.89	0.96
	(0.59)	(0.58)	(0.57)	(0.57)	(0.56)	(0.55)	(0.55)
0.3	0.25	0.62	0.89	0.98	1.00	1.00	1.00
	(0.57)	(0.56)	(0.55)	(0.55)	(0.54)	(0.53)	(0.53)
0.5	0.34	0.72	0.93	0.99	1.00	1.00	1.00
	(0.57)	(0.56)	(0.55)	(0.55)	(0.54)	(0.54)	(0.53)

[a] Risk allele frequency
Population Prevalence of Disease (K) = 0.10
Number of markers on genome-wide platform (M) = 500,000
Overall type I error (α) = 0.05

TABLE 2.3: Statistical power calculations accounting for poor ethnic matching between study cases and public controls. Calculations are for one- and two-stage study designs including study controls (n = 2,000), public controls (n = 5,000) or both. Calculations assume 2,000 cases, M = 500,000 markers in stage 1, a multiplicative genetic model with susceptibility allele frequency = 0.3, K = 0.10 and GRR = 1.3. Power calculated for a range of effective sample-size reductions in public controls due to poor ancestry matching; proportion of cases genotyped in stage 1 analyses of two-stage replication design based on optimized value obtained assuming (*a priori*) that all 5,000 public controls are ethnically matched to study cases.

Percent Reduction in Public Controls	Effective Sample Size Public Controls	Study Controls Only (1-Stage)	Public Controls Only (1-Stage)	2-Stage Replication Design Follow-up Platform (Number of SNPs) 2-Sided test/1-Sided Test in Stage 2					Study + Public Controls (1-Stage)
				100	375	1500	7500	16,500	
0%	5,000	0.78	0.90	0.87/0.89	0.86/0.88	0.84/0.87	0.82/0.85	0.80/0.84	0.96
1%	4,950	0.78	0.90	0.87/0.89	0.86/0.88	0.84/0.87	0.82/0.85	0.80/0.84	0.96
5%	4,750	0.78	0.89	0.86/0.89	0.85/0.88	0.84/0.87	0.82/0.85	0.80/0.84	0.96
10%	4,500	0.78	0.88	0.86/0.88	0.85/0.88	0.84/0.86	0.82/0.85	0.80/0.83	0.96
20%	4,000	0.78	0.86	0.85/0.87	0.84/0.87	0.83/0.86	0.81/0.84	0.80/0.83	0.96
30%	3,500	0.78	0.82	0.84/0.86	0.84/0.86	0.83/0.85	0.81/0.84	0.80/0.83	0.96
40%	3,000	0.78	0.76	0.82/0.85	0.82/0.85	0.82/0.85	0.80/0.83	0.79/0.83	0.95
50%	2,500	0.78	0.68	0.80/0.82	0.80/0.83	0.80/0.83	0.80/0.83	0.79/0.82	0.94
75%	1,250	0.78	0.30	0.63/0.65	0.67/0.70	0.71/0.74	0.74/0.77	0.75/0.78	0.90
90%	500	0.78	0.03	0.27/0.28	0.36/0.38	0.46/0.48	0.57/0.60	0.62/0.65	0.84

TABLE 2.4: Statistical power calculations for two-stage replication design accounting for batch genotype effects between study cases and public controls. Calculations assume 2,000 study cases (spread across stages 1 and 2), 5,000 public controls (stage 1), 2,000 public controls (stage 2) and M = 500,000 markers in stage 1. Power calculated for a multiplicative genetic model with susceptibility minor allele frequency = 0.3, K = 0.10 and GRR = 1.3 across a range of alternative significance thresholds in stage 1 due to batch genotype effects. The proportion of cases genotyped in stage 1 of the two-stage replication design is based on the optimized value obtained assuming (*a priori*) that there are no batch effects (i.e. significance threshold in stage 1 = π_{markers}).

Significance Threshold in Stage 1 After Accounting for Batch Genotype Effects	2-Stage Replication Design Follow-up Platform (Number of SNPs) 2-Sided test/1-Sided Test in Stage 2				
	100	375	1500	7500	16,500
π_{markers} *	0.87/0.89	0.86/0.88	0.84/0.87	0.82/0.85	0.80/0.84
0.99 x π_{markers}	0.87/0.89	0.86/0.88	0.84/0.87	0.82/0.85	0.80/0.84
0.95 x π_{markers}	0.86/0.89	0.85/0.88	0.84/0.87	0.82/0.85	0.80/0.83
0.90 x π_{markers}	0.86/0.89	0.85/0.88	0.84/0.87	0.81/0.85	0.80/0.83
0.80 x π_{markers}	0.86/0.88	0.85/0.87	0.83/0.86	0.81/0.84	0.80/0.83
0.70 x π_{markers}	0.86/0.88	0.85/0.87	0.83/0.86	0.81/0.84	0.79/0.83
0.60 x π_{markers}	0.85/0.87	0.84/0.87	0.82/0.85	0.80/0.83	0.79/0.82
0.50 x π_{markers}	0.84/0.87	0.83/0.86	0.82/0.85	0.79/0.83	0.78/0.81
0.25 x π_{markers}	0.82/0.84	0.81/0.83	0.79/0.82	0.76/0.80	0.75/0.78
0.10 x π_{markers}	0.77/0.80	0.76/0.79	0.74/0.78	0.71/0.75	0.69/0.73

* π_{markers} = 500,000 / # SNPs on follow-up stage 2 genotyping platform (i.e. assuming no batch effects)

TABLE 2.5: Estimated relative cost* (power/proportion of total study samples genotyped in stage 1) of GWA study (M = 500,000 SNPs) for one- and two-stage study designs that include only study controls (n = 2,000), only public controls (n = 5,000) or both. Relative cost estimates assume 2,000 cases, a multiplicative genetic model with susceptibility minor allele frequency = 0.3, K = 0.10 and GRR = 1.3. The relative costs of genotyping 16,000, 7,500, 1,500, and 100 SNPs was assumed to be 1/2, 1/3, 1/5, and 1/12 of the cost of genotyping all 500,000 SNPs on GWA panel, respectively.

		Study Controls Only[1]	Public Controls Only[2]	Public + Study Controls[4]	
One–Stage Genotype Design (All Samples Genotyped on GWA Panel)		\$1.00 (0.78 / 1.00)	\$0.50 (0.90 / 1.00)	\$1.00 (0.97 / 1.00)	
Two-Stage Genotype Design	Follow-Up Platform	Joint Analysis			Replication-Based (Public + Study Controls)[3]
	100	\$0.725 (0.78 / 0.70)	\$0.317 (0.89 / 0.60)	\$0.395 (0.95 / 0.68)	\$0.335 (0.89 / 0.55)
	1,500	\$0.640 (0.78 / 0.55)	\$0.280 (0.89 / 0.45)	\$0.400 (0.95 / 0.50)	\$0.360 (0.87 / 0.40)
	7,500	\$0.640 (0.78 / 0.46)	\$0.287 (0.89 / 0.36)	\$0.448 (0.95 / 0.39)	\$0.437 (0.85 / 0.31)
	16,500	\$0.700 (0.78 / 0.40)	\$0.328 (0.89 / 0.31)	\$0.585 (0.95 / 0.34)	\$0.568 (0.84 / 0.27)

* Costs are calculated relative to the cost of genotyping 2,000 study cases and 2,000 study controls on GWA marker panel

[1] For two-stage design, proportion of samples genotyped in stage 1 represents proportion of cases and study controls (both included in stage 1 genotyping)

[2] For two-stage design, proportion of samples genotyped in stage 1 represents proportion of cases genotyped in stage 1. No study controls are genotyped in stage 1 or stage 2.

[3] Proportion of samples genotyped in stage 1 represents proportion of cases genotyped in stage 1. All study controls are genotyped in stage 2.

[4] For two-stage design, proportion of samples genotyped in stage 1 represents proportion of cases genotyped in stage 1. All study controls are genotyped in stage 2.

2.4 Discussion

Large-scale case-control genetic association studies have proven to be successful in identifying genetic polymorphisms associated with human disease. It has become increasingly clear that the majority of common genetic variants associated with most human disease explain, individually, a relatively small amount of the total disease susceptibility. The modest underlying genetic risk from a given susceptibility allele combined with the high toll of multiple testing inherent with contemporary genotyping platforms necessitates large sample sizes to achieve sufficient statistical power to detect an association. Unfortunately, the sample sizes of todays genetic association studies are constrained by the high cost of genotyping and sample collection. One mechanism that can increase power in genetic association studies of dichotomous traits is to include additional control samples from other studies. Genome-wide genotype data from many different populations are becoming increasingly freely available to scientific researchers through public databases funded by the U.S. National Institutes of Health and other public funding agencies and from private company efforts such as Illuminas iControlDB database. Of note, several recent high-profile GWA studies have included Illuminas iControlDB genotype data demonstrating, empirically, the value of using free public genotype data (Hom et al., 2008; Silverberg et al., 2009; Wrensch et al., 2009). We illustrate, through examples, the gains in statistical power that can be obtained by combining study controls with free public genotype data on unscreened population samples. We also demonstrate that in addition to increasing power, supplementing study control data with free public control genotype data can dramatically decrease overall study cost when utilizing two-stage genotyping designs. This cost reduction is realized due to all study controls being genotyped on the smaller, less expensive, stage 2 genotyping platform and to a smaller proportion of study cases being genotyped in stage 1.

The utilization of free public genotype data is subject to certain limitations and can increase the risk for increased type 1 errors compared to studies that exclude these data. Obviously the biggest potential obstacle of including public control genotype data is the availability of genotype data on a sufficient number of ethnically matched public controls for the same set of SNPs that will be genotyped in study samples. Commercial genotyping platforms are constantly changing, typically adding additional SNPs to established sets of SNPs included on older platforms. This limitation is being mitigated, somewhat, by a more streamlined mechanism for public release of genome-wide SNP data and by collaborations between investigators that study the same or different diseases.

Free public controls typically are not screened for the disease in the current study while study controls often are. It is well known that disease misclassification can reduce statistical power. However, the increasingly large number of free controls that are available to genetic researchers will often overcome this limitation and, as we and others have shown, result in studies with even greater statistical power than studies using a smaller number of screened controls (Edwards et al., 2005; Moskvina et al., 2005; Wellcome Trust Case Control Consortium, 2007; Zheng and Tian, 2005). This benefit is particularly noticeable for traits with low prevalence. It should be noted that in our power calculations, we assumed that free publicly available controls from the general population were not screened for any disease and that screened controls have no disease misclassification. In fact, many public control samples have been ascertained from healthy populations and many disease-screening techniques commonly used to identify controls are not 100% accurate. As a result, the assumptions we used in our power calculations may exaggerate the increased relative power gained by using screened controls when compared to public controls.

A larger concern for utilizing public control genotype data is that observed allele

frequency differences between public controls and study cases may be the consequence of systematic bias due to population stratification or batch effects from differential allele calling between the two samples. Greater differences in background ancestry will likely occur between public controls and cases than between cases and a carefully selected set of controls from the same community. The concern of population stratification can be largely remedied by employing appropriate analytic methods (Price et al., 2006; Roeder and Luca, 2009; Yu et al., 2008), though there still is some concern for a relatively small number of genetic markers under apparent selective pressure. Systematic differences in genotyping calls from plate to plate can also cause bias in genetic association studies (Moskvina et al., 2006; Neale and Purcell, 2008). Despite the availability of public control data from many of the same commercial platforms that would be considered for genotyping sample cases, the inability to account for systematic genotyping errors through experimental design is a source of concern when relying solely on public controls. Unfortunately, DNA is often unavailable on public controls making it difficult to validate, through direct genotyping, any observed differences in genotype frequencies between study cases and public controls. In some circumstances, individual marker fluorescent intensity data from public control samples may be available to facilitate combining these data with the marker data from cases genotyped on the same platform; which would subsequently allow for renormalization and clustering of alleles for the purpose of rescoring genotypes in the combined sample. Further work needs to be done to evaluate the quality-control potential of this approach and, unfortunately, to date these kinds of extensive data on public controls are not routinely available. Given this limitation, the utilization of stringent quality control and including common controls (e.g. HapMap samples) that are present in the public control dataset when genotyping study samples can be critical for identifying individual problematic SNPs. The availability of multiple public control datasets should also aide, through compar-

ison of genotype frequencies among different control populations, in the identification of SNPs that appear to be subject to batch genotype effects.

Public control datasets also typically lack valuable environmental exposure data that is critical to understanding the interplay between genes and environment on the etiology of disease. Even when environmental exposure data have been collected in public controls, it is often not collected or scored in the same manner as in the case study sample, thereby reducing the effectiveness of these data. Furthermore, public controls and cases will usually come from different communities with each community having its own unique set of unmeasured risk factors. These limitations substantially reduce the ability of investigators to evaluate gene-by-environment interactions that are increasingly thought to play a central role in genetic susceptibility.

Recent results from several GWA studies that have included public control genotype data on Caucasian samples have revealed little evidence of strong systematic differences in allele frequencies between previously genotyped public controls and study samples (Hom et al., 2008; Luca et al., 2008; Silverberg et al., 2009; Wrensch et al., 2009; Yu et al., 2008). While the results from these studies are encouraging with respect to control of the overall type I error rate when using public controls, any single result based on public control data should be viewed with some degree of skepticism. It is plausible, given the high quality of genotyping on modern commercial panels, that many SNPs are not subject to strong batch effects when genotyped at different times on the same or different genotyping platforms, but it is very likely that some SNPs are. The recent study by the WTCCC found highly significant differences in allele frequencies for a small number of loci between samples of Caucasians from different communities in Great Britain (Wellcome Trust Case Control Consortium, 2007). The differences were attributed to natural selection, reflecting the historical settlement of ancestors in these different communities from different parts of Europe. It is difficult to control for

the effects of selection using modern analytic methods, such as principal components, when the number of loci under such pressure is small. In the WTCCC study, study cases, like the controls, were largely ascertained across Great Britain, thus substantially reducing the potential impact of bias due to selection. Many studies, however, that will use public control data will include cases and public controls that have been selected from entirely different communities or, possibly, different countries.

We have introduced a replication-based two-stage genotyping design, including both public and study controls, that addresses many of the limitations and concerns regarding the use of public controls while still providing increased power and decreased genotyping costs compared to studies that use only study controls. In this design, public controls and a subset of study cases are used to select a reduced list of SNPs for independent association testing between study controls and the remaining study cases. By this design, the final assessment of whether a SNP is associated with the disease outcome is based entirely on genotype results from study controls, which will be presumably selected from the same community and genotyped at the same time and on the same platform as the set of study cases that have been included in stage 2.

In our power calculations, we have attempted to address the impact of poorly matched (with respect to genetic ancestry) public controls and systematic differences in allele calls for a subset of SNPs under consideration. We have shown, under certain assumptions, that the effect of poorly matched public controls, with respect to ancestry, can have a major impact on studies limited to public controls. However, our proposed two-stage study design, which uses public controls in stage 1 and study controls in stage 2, appears to be relatively robust to this problem even when it is not accounted for in the initial study design. In addition, we have also shown that the proposed two-stage design is robust to batch genotype effects when the SNP associated with disease outcome is not subject to batch genotype effects (or under selection pressure). Should

the SNP associated with the outcome also be subject to batch effects then the impact on power would depend on the direction of the batch effects. Should the direction of the batch effects be the same as the true effect (under the alternative) then power would be increased in stage 1 and the associated SNP would even be more likely to be included in stage 2 genotyping efforts. On the other hand, should the direction of the batch effects be in the opposite direction then power would be decreased in stage 1 and the SNP would be less likely to be included in stage 2 genotyping. Extensive simulations showing the impact of directional batch effects on the SNP associated with disease warrants further future consideration.

The proposed two-stage replication-based approach ensures that the final main-effects analyses, based on stage 2 samples, will be able to incorporate critical environmental exposure data that has been collected from cases and study controls. Furthermore, all study controls and all study cases will have genotype data available on all SNPs that demonstrate a modest degree of evidence of main effects in stage 1, making the study of gene-by-environment interactions feasible for all SNPs genotyped in stage 2 using the complete set of study samples. It has been shown previously that greater power to detect the most plausible gene-by-environment interactions can be achieved by focusing attention on the reduced number of SNPs that demonstrate some evidence of main effects (Kooperberg and Leblanc, 2008).

Our proposed two-stage replication-based sampling design could be particularly valuable for studies that have collected a limited number of study controls (see Model 2, the 2nd example in Table 2.1) or for collaborative studies where some study sites have only collected case samples while other studies have collected sets of matched cases and controls. Collecting new sets of unrelated controls can be problematic for studies that have not already done so (such as family linkage studies). Provided there is not a lot of heterogeneity between the case samples from the various studies, this mechanism

would facilitate the inclusion of additional cases and potentially thousands of public controls, resulting in increased power and likely decreased cost for the combined study.

We have focused our calculations primarily on a study design for a GWA study based on 500,000 SNPs. We have also shown (see Model 3, the 3rd example in Table 2.1) that using public control genotype data in our proposed two-stage design can significantly increase power for smaller follow-up studies as well. Commercial companies are constantly increasing the number of SNPs included on their panels while the cost, per SNP, is declining. It is expected that the results from the 1,000 Genomes Project could lead to commercial genotyping panels that contain millions of SNPs. As a consequence, the examples for which we report power calculations may not reflect study designs in the not too distant future. However, our underlying conclusion that the proposed two-stage genotyping design utilizing public controls can increase statistical power to detect an association and decrease overall study cost while preserving many of the advantages of a well-matched case-control design should hold for future study designs that include more SNPs. We have R software code that is available for investigators who would like to calculate power and make the comparisons for their own studies.

2.5 Supplemental Methods

2.5.1 Explicit Cell Probabilities of the Case-Control Contingency Table

The cell probabilities of the case-control contingency table for the cases were Pr(dd | case), Pr(Dd | case), and Pr(DD | case) and for the controls were Pr(dd | control), Pr(Dd | control), and Pr(DD | control), where d and D were the alleles at a bi-allelic locus and dd, Dd, and DD were the genotype possibilities. The allele frequencies of

d and D were f_d and f_D, respectively, and we assumed Hardy-Weinberg Equilibrium such that the dd, Dd, and DD genotype frequencies were $f_{dd} = f_d^2$, $f_{Dd} = 2f_D f_d$, and $f_{DD} = f_D^2$. The disease prevalence, K, was defined to be

$$K = \Pr(case \mid dd) f_{dd} + \Pr(case \mid Dd) f_{Dd} + \Pr(case \mid DD) f_{DD} \tag{2.2}$$

Multiplicative Genetic Mode-of-Inheritance Risk Model

The genetic relative risk (GRR) under a multiplicative genetic mode-of-inheritance risk model was defined to be

$$GRR = \frac{\Pr(case \mid DD)}{\Pr(case \mid Dd)} = \frac{\Pr(case \mid Dd)}{\Pr(case \mid dd)} \tag{2.3}$$

We had two equations and sought to determine the case and control cell probabilities of the contingency table as described above. From Equation 2.3 followed

$$\Pr(case \mid Dd) = \frac{\Pr(case \mid DD)}{GRR} \text{ and } \Pr(case \mid dd) = \frac{\Pr(case \mid DD)}{GRR^2} \tag{2.4}$$

which we substituted into Equation 2.2 to yield

$$K = \frac{\Pr(case \mid DD)}{GRR^2} f_{dd} + \frac{\Pr(case \mid DD)}{GRR} f_{Dd} + \Pr(case \mid DD) f_{DD}$$

$$K = \Pr(case \mid DD) \cdot \left(\frac{f_{dd}}{GRR^2} + \frac{f_{Dd}}{GRR} + f_{DD} \right) \tag{2.5}$$

Thus,

$$\Pr(case \mid DD) = \frac{K}{\frac{f_{dd}}{GRR^2} + \frac{f_{Dd}}{GRR} + f_{DD}}$$

$$\Pr(case \mid Dd) = \frac{K}{\frac{f_{dd}}{GRR} + f_{Dd} + GRR \cdot f_{DD}}$$

$$\Pr(case \mid dd) = \frac{K}{f_{dd} + GRR \cdot f_{Dd} + GRR^2 \cdot f_{DD}} \tag{2.6}$$

Applying Bayes' Law and then substituting the above penetrances gave us the cell probabilities for the cases,

$$\Pr(DD \mid case) = \frac{\Pr(case \mid DD) \cdot f_{DD}}{K} = \frac{f_{DD}}{\frac{f_{dd}}{GRR^2} + \frac{f_{Dd}}{GRR} + f_{DD}}$$

$$\Pr(Dd \mid case) = \frac{\Pr(case \mid Dd) \cdot f_{Dd}}{K} = \frac{f_{Dd}}{\frac{f_{dd}}{GRR} + f_{Dd} + GRR \cdot f_{DD}}$$

$$\Pr(dd \mid case) = \frac{\Pr(case \mid dd) \cdot f_{dd}}{K} = \frac{f_{dd}}{f_{dd} + GRR \cdot f_{Dd} + GRR^2 \cdot f_{DD}} \tag{2.7}$$

For the study controls, the penetrances followed from the cases' penetrances,

$$\Pr(control \mid DD) = 1 - \Pr(case \mid DD) = 1 - \frac{K}{\frac{f_{dd}}{GRR^2} + \frac{f_{Dd}}{GRR} + f_{DD}}$$

$$\Pr(control \mid Dd) = 1 - \Pr(case \mid Dd) = 1 - \frac{K}{\frac{f_{dd}}{GRR} + f_{Dd} + GRR \cdot f_{DD}}$$

$$\Pr(control \mid dd) = 1 - \Pr(case \mid dd) = 1 - \frac{K}{f_{dd} + GRR \cdot f_{Dd} + GRR^2 \cdot f_{DD}} \quad (2.8)$$

Applying Bayes' Law and then substituting the study controls' penetrances gave us the cell probabilities for the study controls, as we similarly did for the cases,

$$\Pr(DD \mid control) = \frac{\Pr(control \mid DD) \cdot f_{DD}}{1 - K} = \frac{\left(1 - \frac{K}{\frac{f_{dd}}{GRR^2} + \frac{f_{Dd}}{GRR} + f_{DD}}\right) \cdot f_{DD}}{1 - K}$$

$$\Pr(Dd \mid control) = \frac{\Pr(control \mid Dd) \cdot f_{Dd}}{1 - K} = \frac{\left(1 - \frac{K}{\frac{f_{dd}}{GRR} + f_{Dd} + GRR \cdot f_{DD}}\right) \cdot f_{Dd}}{1 - K}$$

$$\Pr(dd \mid control) = \frac{\Pr(control \mid dd) \cdot f_{dd}}{1 - K} = \frac{\left(1 - \frac{K}{f_{dd} + GRR \cdot f_{Dd} + GRR^2 \cdot f_{DD}}\right) \cdot f_{dd}}{1 - K}$$

(2.9)

Dominant Genetic Mode-of-Inheritance Risk Model

The GRR under a dominant genetic mode-of-inheritance risk model was defined to be

$$GRR = \frac{\Pr(case \mid DD)}{\Pr(case \mid dd)} = \frac{\Pr(case \mid Dd)}{\Pr(case \mid dd)} \tag{2.10}$$

Similarly to the proof shown above for the multiplicative model, the genotype probabilities for the cases were

$$\begin{aligned}
\Pr(DD \mid case) &= \frac{f_{DD}}{\frac{f_{dd}}{GRR} + f_{Dd} + f_{DD}} \\
\Pr(Dd \mid case) &= \frac{f_{Dd}}{\frac{f_{dd}}{GRR} + f_{Dd} + f_{DD}} \\
\Pr(dd \mid case) &= \frac{f_{dd}}{f_{dd} + GRR \cdot (f_{Dd} + f_{DD})}
\end{aligned} \tag{2.11}$$

and the genotype probabilities for the controls were

$$\Pr(DD \mid control) = \frac{1 - \frac{K}{\frac{f_{dd}}{GRR} + f_{Dd} + f_{DD}} \cdot f_{DD}}{1 - K}$$

$$\Pr(Dd \mid control) = \frac{1 - \frac{K}{\frac{f_{dd}}{GRR} + f_{Dd} + f_{DD}} \cdot f_{Dd}}{1 - K}$$

$$\Pr(dd \mid control) = \frac{1 - \frac{K}{f_{dd} + GRR \cdot (f_{Dd} + f_{DD})} \cdot f_{dd}}{1 - K} \tag{2.12}$$

Recessive Genetic Mode-of-Inheritance Risk Model

The GRR under a recessive genetic mode-of-inheritance risk model was defined to be

$$GRR = \frac{\Pr(case \mid DD)}{\Pr(case \mid dd)} = \frac{\Pr(case \mid DD)}{\Pr(case \mid Dd)} \tag{2.13}$$

Similarly to the proof shown above for the multiplicative model, the genotype probabilities for the cases were

$$\begin{aligned}
\Pr(DD \mid case) &= \frac{f_{DD}}{\frac{f_{dd}+f_{Dd}}{GRR}+f_{DD}} \\
\Pr(Dd \mid case) &= \frac{f_{Dd}}{f_{dd}+f_{Dd}+GRR\cdot f_{DD}} \\
\Pr(dd \mid case) &= \frac{f_{dd}}{f_{dd}+f_{Dd}+GRR\cdot f_{DD}}
\end{aligned} \tag{2.14}$$

and the genotype probabilities for the controls were

$$\begin{aligned}
\Pr(DD \mid control) &= \frac{1-\frac{K}{\frac{f_{dd}+f_{Dd}}{GRR}+f_{DD}}\cdot f_{DD}}{1-K} \\
\Pr(Dd \mid control) &= \frac{1-\frac{K}{f_{dd}+f_{Dd}+GRR\cdot f_{DD}}\cdot f_{Dd}}{1-K} \\
\Pr(dd \mid control) &= \frac{1-\frac{K}{f_{dd}+f_{Dd}+GRR\cdot f_{DD}}\cdot f_{dd}}{1-K}
\end{aligned} \tag{2.15}$$

Unscreened and Public Controls

For unscreened and public controls, the genotype probabilities for controls were set to the genotype probabilities in the general population, namely

$$\begin{aligned} \Pr(DD \mid control) &= f_{DD} \\ \Pr(Dd \mid control) &= f_{Dd} \\ \Pr(dd \mid control) &= f_{dd} \end{aligned} \tag{2.16}$$

Of note was the observation that the cases' genotype probabilities were not a function of K for the multiplicative, dominant, and recessive genetic mode-of-inheritance risk models (as shown in the above sections), whilst the study controls' genotype probabilities were indeed a function of K. This provided justification for the result in Table 2.1 of the manuscript whereby the power for the 1-stage design with "Study Controls Only" varied with varying levels of K, but the power for the 1-stage design with "Public Controls Only" did not vary with varying levels of K (and similarly for the analogous tables in which the true genetic mode-of-inheritance risk models were dominant and recessive).

2.5.2 Alternative 1- and 2-df Tests

The results in the main manuscript were based on the Cochran-Armitage trend test and assuming an underlying multiplicative genetic mode-of-inheritance risk model. In Supplemental Figure 2.2, we present power curves for the one- and two-stage designs using 2,000 cases, 2,000 study controls and 5,000 public controls over a range of GRRs, disease prevalences, and susceptibility allele frequencies. In Supplemental Table 2.6, we calculated the Cochran-Armitage trend test for the three models considered in Table 2.1 in the main manuscript, but under the assumption that study controls were unscreened

for disease. Supplemental Tables 2.7 through 2.12 present analogous results to Tables 2.1 and 2.2 in the main manuscript, though utilizing the general 2-df, dominant 1-df, and recessive 1-df tests under multiplicative, dominant, and recessive models, respectively. Specifically, Supplemental Tables 2.7, 2.8, and 2.9 are the general, dominant, and recessive versions of Table 2.1 in the main manuscript and Supplemental Tables 2.10, 2.11, and 2.12 are analogous to Table 2.2. The single- and two-staged association study designs as described in the methods of the main manuscript were also used for the Supplemental Tables in terms of the number of study cases, study controls, and public controls and the size of the GWA and follow-up genotyping platforms (Models 1, 2, and 3 of the main manuscript). In addition, the same disease prevalences were specified. However, depending on the genetic model, we allowed the risk allele frequency (f_D) and GRR to vary. For the 1- and 2-df tests, we computed power using the "cost effective" (CE) method proposed by Bukszár and van den Oord (Bukszár and van den Oord, 2006a). The CE is an approximation for computing the power of Pearson's statistic for 2 x m (where m refers to the number of categories) contingency tables that is accurate and efficient in terms of computer time. The authors point out (Bukszár and van den Oord, 2006b) that the CE is very close to the true value of the distribution of Pearson's statistic and more accurate than a commonly used approximation (based on a non-central chi-square) that overestimates power in some scenarios and underestimates it in others.

General 2-df Test

For Supplemental Tables 2.7 and 2.10, the general 2-df test was employed assuming an underlying multiplicative genetic mode-of-inheritance risk model. The specific genotype cell probabilities for the cases and controls are shown above. For Supplemental Table 2.7, as in the main manuscript, f_D and GRR were set to 0.3 and 1.3, respectively. In

order to compute power using the general 2-df test, we carried out the CE for 2 x m tables where m = 3 columns / categories (genotypes dd, Dd, and DD) and the rows pertained to the cases and controls, using the R script costeff2by3 provided by Bukszár and van den Oord (http://www.vipbg.vcu.edu/ edwin/). Contrary to the 1-df test, closed form analytical formulae did not exist for the 2 x 3 tables, though numerical solutions were computed with the costeff2by3 R code.

Dominant 1-df Test

For Supplemental Tables 2.8 and 2.11, the dominant 1-df test was employed assuming an underlying dominant genetic mode-of-inheritance risk model. The specific genotype cell probabilities for the cases and controls are shown above. For Supplemental Table 2.8, f_D was set to 0.3 (as with the multiplicative model), though the GRR was set to 1.4. In order to compute power using the dominant 1-df test, we carried out the CE for 2 x m tables where m = 2 columns / categories (genotypes dd and Dd, or DD, i.e. the Dd and DD genotype columns were combined) and the rows pertained to the cases and controls. The power for critical value c (corresponding to the 1 - type I error) of a central chi-square distribution was (Bukszár and van den Oord, 2006a)

$$1 - F_{\chi^2} \quad \frac{c}{\lambda} \tag{2.17}$$

where λ was the largest eigenvalue of matrix J (discussed by Bukszár and van den Oord) and F_{χ^2} was the cdf of the non-central chi-square distribution with 1 degree of freedom and non-centrality parameter

$$\omega = \frac{(p_1 - q_1)^2 pqn}{\lambda(pp_1 + qq_1)(pp_2 + qq_2)} \tag{2.18}$$

where n was the total sample size, p was the proportion of controls in the total sample, q = 1 - p was the proportion of cases in the total sample, and subscripts 1 and 2 referred to the two genotype categories. These computations were carried out with the R script costeff2by2 provided by Bukszár and van den Oord (http://www.vipbg.vcu.edu/ edwin/).

Recessive 1-df Test

For Supplemental Tables 2.9 and 2.12, the recessive 1-df test was employed assuming an underlying recessive genetic mode-of-inheritance risk model. The specific genotype cell probabilities for the cases and controls are shown above. For Supplemental Table 2.9, f_D and GRR were set to 0.5 and 1.45, respectively. The power calculations were performed in the same manner as the dominant 1-df test detailed above though the 2 x 2 table was constructed differently, namely, the dd and Dd columns were merged.

2.6 Supplemental Results

Under a multiplicative, dominant, and recessive genetic mode-of-inheritance risk model and conducting a general 2-df, dominant 1-df , and recessive 1-df test, respectively, the overall observations discussed in the main manuscript pertaining to the performances of the study designs when using the Cochran-Armitage trend test (Table 2.1) were applicable for the alternative tests (Supplemental Tables 2.7, 2.8, and 2.9). For instance, the power of the proposed two-stage replication-based design that used both public and study controls was significantly greater than the power for the single-stage study design that used only study controls for nearly all study designs considered. The single-stage study design based solely on public controls had greater power than the single-stage study design based solely on screened study controls for many alternative models when

the number of public controls was greater than the number of study controls. The two-stage replication-based study design compared favorably to the singe-stage study design that used public controls under most alternative models, particularly for smaller public control samples and higher disease prevalences.

When carrying out a general 2-df test under a multiplicative model, as expected almost all of the study designs across varying levels of disease prevalences resulted in a loss of power (Supplemental Table 2.7), compared to the Cochran-Armitage trend test (Table 2.1).

For the alternative tests assuming multiplicative, dominant, and recessive models, our results showed that the optimal choice of the proportion of cases used in stage 1 were robust across a range of different possible alternative models (Supplemental Tables 2.10, 2.11, and 2.12), which we had also noted in the main manuscript (Table 2.2) for the Cochran-Armitage trend test under a multiplicative model. The other observations discussed in the main manuscript also applied to Supplemental Tables 2.10, 2.11, and 2.12.

In Supplemental Table 2.10, we noted that the powers for our two-stage design across a range of GRRs, f_Ds, and follow-up platforms were lower for the general test as compared to the Cochran-Armitage test (Table 2.2 of the main manuscript). Lastly, despite the genetic inheritance model and test conducted, the proportion of cases in stage 1 across the ranges of GRRs, f_Ds, and follow-up platforms were about the same as the proportions seen in Table 2.2.

TABLE 2.6: Power of the Cochran-Armitage trend test for 1- and 2-stage study designs across a range of sample sizes, SNPs in stage 1, and disease prevalences. Study controls are assumed to be unscreened for disease and to have the same disease risk as the general population. Note, under this assumption, power is constant across different values of disease prevalence for all study designs.

	Number of Public Controls											
	1,000			3,000			5,000			10,000		
Unscreened Controls Only[a]	Public Controls Only[b]	2-Stage[c] (2-sided / 1-sided)	Unscreened + Public Controls[d]	Public Controls Only[b]	2-Stage[c] (2-sided / 1-sided)	Unscreened + Public Controls[d]	Public Controls Only[b]	2-Stage[c] (2-sided / 1-sided)	Unscreened + Public Controls[d]	Public Controls Only[b]	2-Stage[c] (2-sided / 1-sided)	Unscreened + Public Controls[d]
2,000 Cases and 2,000 Unscreened Controls / 500,000 Genome-wide Markers												
0.57	0.19	0.57 / 0.62	0.76	0.76	0.73 / 0.77	0.90	0.90	0.78 / 0.82	0.94	0.97	0.81 / 0.84	0.97
2,000 Cases and 1,000 Unscreened Controls / 500,000 Genome-wide Markers												
0.19	0.19	0.37 / 0.41	0.56	0.76	0.57 / 0.62	0.86	0.90	0.60 / 0.66	0.93	0.97	0.63 / 0.69	0.97
1,250 Cases and 1,250 Unscreened Controls / 7,500 Genome-wide Markers												
0.44	0.34	0.47 / 0.53	0.68	0.76	0.55 / 0.61	0.83	0.86	0.56 / 0.62	0.88	0.92	0.57 / 0.64	0.92

[a] 1-stage design using all unscreened controls
[b] 1-stage design using all public controls
[c] Optimal 2-stage replication design using all public controls in stage 1 and all unscreened controls in stage 2, 2-sided test in both stages / 2-sided test in stage 1 and 1-sided test in stage 2
[d] 1-stage design pooling the unscreened and public controls
Risk allele frequency (f_D) = 0.3, genetic relative risk (GRR) = 1.3 assuming a multiplicative model, overall type I error (α) = 0.05

TABLE 2.7: Power of the general 2-df test for 1- and 2-stage study designs across a range of sample sizes, SNPs in stage 1, and disease prevalences. Study controls are assumed to be screened and disease free.

		Number of Public Controls											
		1,000			3,000			5,000			10,000		
K[a]	Study Controls Only	Public Controls Only	2-Stage[b] (2-sided / 1-sided)	Study + Public Controls	Public Controls Only	2-Stage[b] (2-sided / 1-sided)	Study + Public Controls	Public Controls Only	2-Stage[b] (2-sided / 1-sided)	Study + Public Controls	Public Controls Only	2-Stage[b] (2-sided / 1-sided)	Study + Public Controls
2,000 Cases and 2,000 Screened Study Controls / 500,000 SNPs													
0.0001	0.46	0.12	0.44 / 0.48	0.68	0.68	0.61 / 0.65	0.85	0.85	0.66 / 0.71	0.91	0.94	0.70 / 0.74	0.95
0.01	0.48	0.12	0.46 / 0.50	0.69	0.68	0.62 / 0.66	0.85	0.85	0.67 / 0.72	0.91	0.94	0.71 / 0.75	0.96
0.05	0.58	0.12	0.52 / 0.56	0.75	0.68	0.66 / 0.70	0.88	0.85	0.72 / 0.76	0.93	0.94	0.76 / 0.79	0.96
0.1	0.70	0.12	0.61 / 0.65	0.82	0.68	0.72 / 0.75	0.91	0.85	0.77 / 0.80	0.94	0.94	0.81 / 0.84	0.96
0.25	0.96	0.12	0.83 / 0.85	0.96	0.68	0.91 / 0.93	0.97	0.85	0.92 / 0.94	0.98	0.94	0.93 / 0.94	0.98
0.5	1.00	0.12	0.96 / 0.96	1.00	0.68	1.00 / 1.00	1.00	0.85	1.00 / 1.00	1.00	0.94	1.00 / 1.00	1.00
2,000 Cases and 1,000 Screened Study Controls / 500,000 SNPs													
0.0001	0.12	0.12	0.25 / 0.28	0.47	0.68	0.43 / 0.49	0.78	0.85	0.47 / 0.53	0.88	0.94	0.50 / 0.56	0.95
0.01	0.13	0.12	0.25 / 0.29	0.47	0.68	0.44 / 0.50	0.80	0.85	0.48 / 0.54	0.88	0.94	0.51 / 0.57	0.95
0.05	0.18	0.12	0.29 / 0.33	0.52	0.68	0.49 / 0.55	0.81	0.85	0.54 / 0.59	0.89	0.94	0.57 / 0.63	0.95
0.1	0.26	0.12	0.35 / 0.39	0.59	0.68	0.56 / 0.61	0.84	0.85	0.61 / 0.66	0.91	0.94	0.64 / 0.70	0.96
0.25	0.65	0.12	0.62 / 0.66	0.80	0.68	0.76 / 0.79	0.91	0.85	0.81 / 0.84	0.94	0.94	0.85 / 0.88	0.97
0.5	1.00	0.12	0.95 / 0.95	1.00	0.68	0.99 / 0.99	0.99	0.85	1.00 / 1.00	0.99	0.94	1.00 / 1.00	0.99
1,250 Cases and 1,250 Screened Study Controls / 7,500 SNPs													
0.0001	0.33	0.24	0.34 / 0.39	0.57	0.67	0.41 / 0.47	0.76	0.78	0.43 / 0.49	0.82	0.86	0.44 / 0.50	0.87
0.01	0.35	0.24	0.35 / 0.40	0.59	0.67	0.42 / 0.48	0.76	0.78	0.44 / 0.50	0.81	0.86	0.45 / 0.51	0.87
0.05	0.42	0.24	0.39 / 0.44	0.63	0.67	0.47 / 0.53	0.77	0.78	0.49 / 0.55	0.82	0.86	0.50 / 0.56	0.88
0.1	0.52	0.24	0.46 / 0.52	0.69	0.67	0.54 / 0.60	0.81	0.78	0.56 / 0.62	0.85	0.86	0.58 / 0.63	0.89
0.25	0.86	0.24	0.74 / 0.78	0.87	0.67	0.77 / 0.81	0.89	0.78	0.78 / 0.82	0.90	0.86	0.80 / 0.83	0.91
0.5	1.00	0.24	0.97 / 0.97	1.00	0.67	0.99 / 0.99	0.99	0.78	0.99 / 0.99	0.98	0.86	0.99 / 1.00	0.97

[a] Population prevalence of disease
[b] Optimal 2-stage replication design using all public controls in stage 1 and all screened controls in stage 2, 2-sided test in both stages / 2-sided test in stage 1 and 1-sided test in stage 2
Risk allele frequency (f_D) = 0.3, genetic relative risk (GRR) = 1.3 assuming a multiplicative model, overall type I error (α) = 0.05

TABLE 2.8: Power of the dominant test for 1- and 2-stage study designs across a range of sample sizes, SNPs in stage 1, and disease prevalences. Study controls are assumed to be screened and disease free.

		Number of Public Controls											
		1,000			3,000			5,000			10,000		
K[a]	Study Controls Only	Public Controls Only	2-Stage[b] (2-sided / 1-sided)	Study + Public Controls	Public Controls Only	2-Stage[b] (2-sided / 1-sided)	Study + Public Controls	Public Controls Only	2-Stage[b] (2-sided / 1-sided)	Study + Public Controls	Public Controls Only	2-Stage[b] (2-sided / 1-sided)	Study + Public Controls
2,000 Cases and 2,000 Screened Study Controls / 500,000 SNPs													
0.0001	0.48	0.16	0.48 / 0.53	0.67	0.67	0.64 / 0.68	0.83	0.84	0.69 / 0.73	0.89	0.93	0.72 / 0.77	0.94
0.01	0.50	0.16	0.50 / 0.54	0.67	0.67	0.65 / 0.69	0.83	0.84	0.70 / 0.74	0.89	0.93	0.73 / 0.77	0.94
0.05	0.59	0.16	0.56 / 0.60	0.74	0.67	0.69 / 0.73	0.86	0.84	0.74 / 0.78	0.90	0.93	0.77 / 0.81	0.95
0.1	0.70	0.16	0.64 / 0.68	0.80	0.67	0.74 / 0.77	0.89	0.84	0.79 / 0.82	0.92	0.93	0.82 / 0.85	0.95
0.25	0.96	0.16	0.85 / 0.86	0.96	0.67	0.91 / 0.93	0.96	0.84	0.92 / 0.94	0.97	0.93	0.93 / 0.94	0.97
0.5	1.00	0.16	0.97 / 0.97	1.00	0.67	1.00 / 1.00	1.00	0.84	1.00 / 1.00	1.00	0.93	1.00 / 1.00	1.00
2,000 Cases and 1,000 Screened Study Controls / 500,000 SNPs													
0.0001	0.16	0.16	0.30 / 0.34	0.47	0.67	0.48 / 0.54	0.76	0.84	0.52 / 0.58	0.87	0.93	0.55 / 0.61	0.93
0.01	0.17	0.16	0.31 / 0.35	0.49	0.67	0.49 / 0.55	0.77	0.84	0.53 / 0.59	0.86	0.93	0.56 / 0.62	0.94
0.05	0.22	0.16	0.35 / 0.39	0.53	0.67	0.54 / 0.59	0.80	0.84	0.58 / 0.64	0.88	0.93	0.61 / 0.67	0.94
0.1	0.30	0.16	0.41 / 0.45	0.58	0.67	0.60 / 0.65	0.81	0.84	0.65 / 0.70	0.89	0.93	0.68 / 0.73	0.94
0.25	0.66	0.16	0.66 / 0.70	0.79	0.67	0.78 / 0.81	0.90	0.84	0.83 / 0.85	0.93	0.93	0.86 / 0.88	0.96
0.5	1.00	0.16	0.95 / 0.96	1.00	0.67	0.99 / 0.99	0.99	0.84	0.99 / 1.00	0.98	0.93	1.00 / 1.00	0.98
1,250 Cases and 1,250 Screened Study Controls / 7,500 SNPs													
0.0001	0.37	0.29	0.39 / 0.45	0.58	0.67	0.46 / 0.52	0.75	0.78	0.48 / 0.54	0.80	0.85	0.49 / 0.55	0.85
0.01	0.39	0.29	0.40 / 0.46	0.58	0.67	0.47 / 0.53	0.73	0.78	0.49 / 0.55	0.80	0.85	0.50 / 0.56	0.85
0.05	0.46	0.29	0.45 / 0.50	0.63	0.67	0.52 / 0.58	0.76	0.78	0.54 / 0.60	0.82	0.85	0.55 / 0.61	0.86
0.1	0.55	0.29	0.51 / 0.56	0.69	0.67	0.59 / 0.64	0.79	0.78	0.60 / 0.66	0.84	0.85	0.62 / 0.67	0.87
0.25	0.85	0.29	0.76 / 0.80	0.86	0.67	0.79 / 0.82	0.88	0.78	0.80 / 0.83	0.89	0.85	0.81 / 0.84	0.90
0.5	1.00	0.29	0.97 / 0.98	1.00	0.67	0.99 / 0.99	0.99	0.78	0.99 / 0.99	0.97	0.85	0.99 / 1.00	0.95

[a] Population prevalence of disease

[b] Optimal 2-stage replication design using all public controls in stage 1 and all screened controls in stage 2, 2-sided test in both stages / 2-sided test in stage 1 and 1-sided test in stage 2

Risk allele frequency (f_D) = 0.3, genetic relative risk (GRR) = 1.4 assuming a dominant model, overall type I error (α) = 0.05

TABLE 2.9: Power of the recessive test for 1- and 2-stage study designs across a range of sample sizes, SNPs in stage 1, and disease prevalences. Study controls are assumed to be screened and disease free.

		Number of Public Controls											
		1,000			3,000			5,000			10,000		
K[a]	Study Controls Only	Public Controls Only	2-Stage[b] (2-sided / 1-sided)	Study + Public Controls	Public Controls Only	2-Stage[b] (2-sided / 1-sided)	Study + Public Controls	Public Controls Only	2-Stage[b] (2-sided / 1-sided)	Study + Public Controls	Public Controls Only	2-Stage[b] (2-sided / 1-sided)	Study + Public Controls
2,000 Cases and 2,000 Screened Study Controls / 500,000 SNPs													
0.0001	0.49	0.14	0.50 / 0.55	0.68	0.70	0.67 / 0.71	0.86	0.86	0.72 / 0.76	0.91	0.95	0.75 / 0.79	0.96
0.01	0.51	0.14	0.52 / 0.56	0.70	0.70	0.68 / 0.72	0.86	0.86	0.73 / 0.77	0.91	0.95	0.76 / 0.80	0.96
0.05	0.60	0.14	0.58 / 0.63	0.76	0.70	0.72 / 0.76	0.89	0.86	0.77 / 0.80	0.92	0.95	0.80 / 0.84	0.96
0.1	0.72	0.14	0.67 / 0.70	0.83	0.70	0.77 / 0.80	0.91	0.86	0.82 / 0.84	0.94	0.95	0.85 / 0.87	0.97
0.25	0.97	0.14	0.87 / 0.88	0.97	0.70	0.93 / 0.94	0.97	0.86	0.94 / 0.95	0.98	0.95	0.95 / 0.96	0.98
0.5	1.00	0.14	0.97 / 0.97	1.00	0.70	1.00 / 1.00	1.00	0.86	1.00 / 1.00	1.00	0.95	1.00 / 1.00	1.00
2,000 Cases and 1,000 Screened Study Controls / 500,000 SNPs													
0.0001	0.14	0.14	0.31 / 0.35	0.48	0.70	0.50 / 0.56	0.79	0.86	0.54 / 0.60	0.89	0.95	0.56 / 0.63	0.95
0.01	0.15	0.14	0.32 / 0.36	0.49	0.70	0.51 / 0.57	0.80	0.86	0.55 / 0.61	0.89	0.95	0.58 / 0.64	0.95
0.05	0.20	0.14	0.36 / 0.40	0.54	0.70	0.56 / 0.62	0.82	0.86	0.60 / 0.66	0.90	0.95	0.63 / 0.69	0.96
0.1	0.28	0.14	0.42 / 0.46	0.60	0.70	0.63 / 0.68	0.83	0.86	0.67 / 0.72	0.91	0.95	0.70 / 0.75	0.96
0.25	0.67	0.14	0.68 / 0.72	0.81	0.70	0.81 / 0.83	0.91	0.86	0.85 / 0.88	0.94	0.95	0.88 / 0.91	0.97
0.5	1.00	0.14	0.96 / 0.97	1.00	0.70	1.00 / 1.00	0.99	0.86	1.00 / 1.00	0.99	0.95	1.00 / 1.00	0.99
1,250 Cases and 1,250 Screened Study Controls / 7,500 SNPs													
0.0001	0.38	0.28	0.42 / 0.47	0.60	0.71	0.49 / 0.55	0.78	0.81	0.50 / 0.56	0.84	0.89	0.51 / 0.57	0.89
0.01	0.39	0.28	0.43 / 0.48	0.61	0.71	0.50 / 0.56	0.78	0.81	0.52 / 0.58	0.84	0.89	0.53 / 0.59	0.89
0.05	0.47	0.28	0.47 / 0.52	0.65	0.71	0.55 / 0.61	0.80	0.81	0.57 / 0.63	0.85	0.89	0.58 / 0.64	0.90
0.1	0.57	0.28	0.54 / 0.59	0.71	0.71	0.62 / 0.67	0.83	0.81	0.64 / 0.69	0.86	0.89	0.65 / 0.70	0.90
0.25	0.88	0.28	0.79 / 0.82	0.89	0.71	0.82 / 0.85	0.90	0.81	0.83 / 0.86	0.91	0.89	0.84 / 0.87	0.92
0.5	1.00	0.28	0.98 / 0.98	1.00	0.71	0.99 / 1.00	0.99	0.81	1.00 / 1.00	0.98	0.89	1.00 / 1.00	0.97

[a] Population prevalence of disease

[b] Optimal 2-stage replication design using all public controls in stage 1 and all screened controls in stage 2, 2-sided test in both stages / 2-sided test in stage 1 and 1-sided test in stage 2

Risk allele frequency (f_D) = 0.5, genetic relative risk (GRR) = 1.45 assuming a recessive model, overall type I error (α) = 0.05

TABLE 2.10: Power for the general 2-df test and the proportion of cases in stage 1 that optimizes power (in parenthesis) in a two-stage replication-based GWA study with 2,000 Cases / 5,000 public controls (stage 1) / 2,000 screened controls (stage 2), assuming a multiplicative model.

	Genetic Relative Risk										
f_D[a]	1.25	1.275	1.3	1.325	1.35	1.375	1.4	1.425	1.45	1.475	1.5
Follow-up Platform: 16500											
0.1	0.02	0.05	0.09	0.15	0.23	0.33	0.45	0.56	0.67	0.77	0.84
	(0.31)	(0.30)	(0.29)	(0.29)	(0.28)	(0.28)	(0.27)	(0.27)	(0.27)	(0.27)	(0.27)
0.3	0.35	0.53	0.70	0.83	0.91	0.96	0.98	0.99	1.00	1.00	1.00
	(0.28)	(0.28)	(0.27)	(0.28)	(0.28)	(0.28)	(0.29)	(0.30)	(0.30)	(0.31)	(0.31)
0.5	0.47	0.65	0.80	0.90	0.96	0.98	0.99	1.00	1.00	1.00	1.00
	(0.28)	(0.28)	(0.28)	(0.28)	(0.28)	(0.29)	(0.29)	(0.30)	(0.30)	(0.31)	(0.31)
Follow-up Platform: 7500											
0.1	0.03	0.05	0.10	0.16	0.25	0.35	0.47	0.58	0.69	0.78	0.85
	(0.35)	(0.35)	(0.34)	(0.33)	(0.32)	(0.32)	(0.32)	(0.31)	(0.31)	(0.31)	(0.31)
0.3	0.37	0.55	0.72	0.84	0.92	0.97	0.99	1.00	1.00	1.00	1.00
	(0.32)	(0.32)	(0.32)	(0.32)	(0.32)	(0.32)	(0.33)	(0.33)	(0.34)	(0.34)	(0.35)
0.5	0.48	0.67	0.81	0.91	0.96	0.98	0.99	1.00	1.00	1.00	1.00
	(0.32)	(0.32)	(0.32)	(0.32)	(0.32)	(0.33)	(0.33)	(0.33)	(0.34)	(0.34)	(0.35)
Follow-up Platform: 1500											
0.1	0.03	0.06	0.11	0.18	0.28	0.39	0.51	0.62	0.72	0.81	0.87
	(0.44)	(0.43)	(0.42)	(0.42)	(0.41)	(0.40)	(0.40)	(0.40)	(0.40)	(0.40)	(0.40)
0.3	0.40	0.58	0.74	0.86	0.93	0.97	0.99	1.00	1.00	1.00	1.00
	(0.40)	(0.40)	(0.40)	(0.40)	(0.40)	(0.40)	(0.40)	(0.40)	(0.41)	(0.41)	(0.41)
0.5	0.51	0.69	0.83	0.92	0.96	0.99	1.00	1.00	1.00	1.00	1.00
	(0.40)	(0.40)	(0.40)	(0.40)	(0.40)	(0.40)	(0.40)	(0.40)	(0.41)	(0.41)	(0.41)
Follow-up Platform: 375											
0.1	0.03	0.07	0.12	0.20	0.30	0.41	0.53	0.65	0.75	0.83	0.89
	(0.51)	(0.50)	(0.49)	(0.49)	(0.48)	(0.48)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)
0.3	0.42	0.60	0.76	0.87	0.94	0.97	0.99	1.00	1.00	1.00	1.00
	(0.48)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)
0.5	0.53	0.71	0.84	0.92	0.97	0.99	1.00	1.00	1.00	1.00	1.00
	(0.47)	(0.47)	(0.47)	(0.47)	(0.46)	(0.46)	(0.46)	(0.46)	(0.47)	(0.47)	(0.47)
Follow-up Platform: 100											
0.1	0.04	0.07	0.13	0.21	0.32	0.43	0.55	0.66	0.76	0.84	0.90
	(0.57)	(0.56)	(0.56)	(0.55)	(0.55)	(0.54)	(0.54)	(0.54)	(0.54)	(0.53)	(0.53)
0.3	0.44	0.62	0.77	0.88	0.94	0.98	0.99	1.00	1.00	1.00	1.00
	(0.54)	(0.54)	(0.54)	(0.53)	(0.53)	(0.53)	(0.53)	(0.52)	(0.52)	(0.52)	(0.52)
0.5	0.55	0.72	0.85	0.93	0.97	0.99	1.00	1.00	1.00	1.00	1.00
	(0.54)	(0.54)	(0.53)	(0.53)	(0.53)	(0.53)	(0.52)	(0.52)	(0.52)	(0.52)	(0.52)

[a] Risk allele frequency
Population Prevalence of Disease (K) = 0.10
Number of markers on genome-wide platform (M) = 500,000
Overall type I error (α) = 0.05

TABLE 2.11: Power for the dominant test and the proportion of cases in stage 1 that optimizes power (in parenthesis) in a two-stage replication-based GWA study with 2,000 Cases / 5,000 public controls (stage 1) / 2,000 screened controls (stage 2), assuming a dominant model.

	Genetic Relative Risk												
f_D[a]	1.35	1.375	1.4	1.425	1.45	1.475	1.5	1.525	1.55	1.575	1.6	1.625	1.65
Follow-up Platform: 16500													
0.1	0.19	0.28	0.37	0.48	0.58	0.67	0.76	0.83	0.88	0.92	0.95	0.97	0.98
	(0.28)	(0.28)	(0.27)	(0.27)	(0.27)	(0.27)	(0.27)	(0.27)	(0.27)	(0.27)	(0.28)	(0.29)	(0.28)
0.3	0.49	0.61	0.72	0.81	0.87	0.92	0.95	0.97	0.99	0.99	1.00	1.00	1.00
	(0.26)	(0.26)	(0.26)	(0.26)	(0.26)	(0.27)	(0.27)	(0.27)	(0.27)	(0.28)	(0.28)	(0.29)	(0.29)
0.5	0.17	0.25	0.33	0.42	0.51	0.60	0.68	0.76	0.82	0.87	0.90	0.93	0.95
	(0.27)	(0.27)	(0.26)	(0.26)	(0.26)	(0.26)	(0.26)	(0.26)	(0.26)	(0.26)	(0.26)	(0.26)	(0.26)
Follow-up Platform: 7500													
0.1	0.21	0.29	0.39	0.50	0.60	0.69	0.77	0.84	0.89	0.93	0.95	0.97	0.98
	(0.32)	(0.32)	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)	(0.32)	(0.32)
0.3	0.50	0.63	0.73	0.82	0.88	0.93	0.96	0.98	0.99	0.99	1.00	1.00	1.00
	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)	(0.32)	(0.32)	(0.32)	(0.33)
0.5	0.18	0.26	0.34	0.43	0.53	0.62	0.70	0.77	0.83	0.87	0.91	0.94	0.96
	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)	(0.30)	(0.31)	(0.31)	(0.30)	(0.30)	(0.30)	(0.30)	(0.30)
Follow-up Platform: 1500													
0.1	0.23	0.33	0.43	0.54	0.64	0.73	0.80	0.86	0.91	0.94	0.96	0.98	0.99
	(0.41)	(0.40)	(0.40)	(0.40)	(0.40)	(0.39)	(0.39)	(0.39)	(0.39)	(0.39)	(0.40)	(0.40)	(0.40)
0.3	0.54	0.66	0.76	0.84	0.90	0.94	0.96	0.98	0.99	0.99	1.00	1.00	1.00
	(0.40)	(0.39)	(0.39)	(0.39)	(0.39)	(0.39)	(0.39)	(0.39)	(0.39)	(0.39)	(0.40)	(0.40)	(0.40)
0.5	0.20	0.28	0.37	0.46	0.55	0.64	0.72	0.79	0.84	0.89	0.92	0.95	0.96
	(0.40)	(0.40)	(0.40)	(0.40)	(0.39)	(0.39)	(0.39)	(0.39)	(0.39)	(0.39)	(0.39)	(0.39)	(0.39)
Follow-up Platform: 375													
0.1	0.25	0.35	0.46	0.56	0.66	0.75	0.82	0.88	0.92	0.95	0.97	0.98	0.99
	(0.48)	(0.48)	(0.48)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.46)	(0.46)	(0.46)	(0.46)	(0.47)
0.3	0.56	0.68	0.78	0.85	0.91	0.95	0.97	0.98	0.99	1.00	1.00	1.00	1.00
	(0.47)	(0.47)	(0.47)	(0.47)	(0.46)	(0.46)	(0.46)	(0.46)	(0.46)	(0.46)	(0.46)	(0.46)	(0.46)
0.5	0.21	0.29	0.38	0.48	0.57	0.66	0.74	0.80	0.86	0.90	0.93	0.95	0.97
	(0.48)	(0.48)	(0.48)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.46)	(0.46)	(0.46)	(0.46)	(0.46)
Follow-up Platform: 100													
0.1	0.27	0.37	0.47	0.58	0.68	0.76	0.83	0.88	0.92	0.95	0.97	0.98	0.99
	(0.55)	(0.55)	(0.55)	(0.54)	(0.54)	(0.54)	(0.54)	(0.53)	(0.53)	(0.53)	(0.53)	(0.53)	(0.53)
0.3	0.57	0.69	0.79	0.86	0.91	0.95	0.97	0.98	0.99	1.00	1.00	1.00	1.00
	(0.54)	(0.54)	(0.54)	(0.54)	(0.53)	(0.53)	(0.53)	(0.53)	(0.53)	(0.52)	(0.52)	(0.52)	(0.52)
0.5	0.22	0.30	0.39	0.49	0.59	0.67	0.75	0.81	0.86	0.90	0.93	0.95	0.97
	(0.55)	(0.55)	(0.55)	(0.54)	(0.54)	(0.54)	(0.54)	(0.54)	(0.54)	(0.53)	(0.53)	(0.53)	(0.53)

[a] Risk allele frequency
Population Prevalence of Disease (K) = 0.10
Number of markers on genome-wide platform (M) = 500,000
Overall type I error (α) = 0.05

TABLE 2.12: Power for the recessive test and the proportion of cases in stage 1 that optimizes power (in parenthesis) in a two-stage replication-based GWA study with 2,000 Cases / 5,000 public controls (stage 1) / 2,000 screened controls (stage 2), assuming a recessive model.

	Genetic Relative Risk												
f_D[a]	1.4	1.425	1.45	1.475	1.5	1.525	1.55	1.575	1.6	1.625	1.65	1.675	1.7
Follow-up Platform: 16500													
0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
0.3	0.06	0.09	0.13	0.18	0.23	0.30	0.37	0.44	0.51	0.59	0.66	0.72	0.78
	(0.30)	(0.30)	(0.29)	(0.29)	(0.28)	(0.28)	(0.28)	(0.27)	(0.27)	(0.27)	(0.27)	(0.27)	(0.27)
0.5	0.54	0.65	0.74	0.82	0.88	0.93	0.95	0.97	0.99	0.99	1.00	1.00	1.00
	(0.27)	(0.27)	(0.27)	(0.27)	(0.27)	(0.27)	(0.27)	(0.28)	(0.28)	(0.29)	(0.29)	(0.29)	(0.30)
Follow-up Platform: 7500													
0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
0.3	0.07	0.10	0.14	0.19	0.25	0.32	0.39	0.46	0.54	0.61	0.68	0.74	0.79
	(0.35)	(0.34)	(0.33)	(0.33)	(0.33)	(0.32)	(0.32)	(0.32)	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)
0.5	0.56	0.67	0.76	0.83	0.89	0.93	0.96	0.98	0.99	0.99	1.00	1.00	1.00
	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)	(0.32)	(0.32)	(0.32)	(0.33)	(0.33)	(0.33)
Follow-up Platform: 1500													
0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
0.3	0.08	0.12	0.17	0.22	0.29	0.36	0.43	0.51	0.58	0.65	0.72	0.77	0.82
	(0.43)	(0.42)	(0.42)	(0.41)	(0.41)	(0.41)	(0.40)	(0.40)	(0.40)	(0.40)	(0.40)	(0.39)	(0.39)
0.5	0.60	0.70	0.79	0.86	0.91	0.94	0.97	0.98	0.99	0.99	1.00	1.00	1.00
	(0.40)	(0.39)	(0.39)	(0.39)	(0.39)	(0.39)	(0.39)	(0.40)	(0.40)	(0.40)	(0.40)	(0.40)	(0.41)
Follow-up Platform: 375													
0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
0.3	0.09	0.13	0.18	0.24	0.31	0.39	0.46	0.54	0.61	0.68	0.74	0.79	0.84
	(0.50)	(0.49)	(0.49)	(0.49)	(0.48)	(0.48)	(0.48)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)
0.5	0.62	0.72	0.80	0.87	0.92	0.95	0.97	0.98	0.99	1.00	1.00	1.00	1.00
	(0.47)	(0.47)	(0.47)	(0.47)	(0.46)	(0.46)	(0.46)	(0.46)	(0.47)	(0.46)	(0.47)	(0.47)	(0.47)
Follow-up Platform: 100													
0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
0.3	0.10	0.14	0.20	0.26	0.33	0.40	0.48	0.56	0.63	0.70	0.76	0.81	0.85
	(0.56)	(0.56)	(0.56)	(0.55)	(0.55)	(0.55)	(0.55)	(0.54)	(0.54)	(0.54)	(0.54)	(0.54)	(0.54)
0.5	0.64	0.74	0.82	0.88	0.92	0.95	0.97	0.98	0.99	1.00	1.00	1.00	1.00
	(0.54)	(0.54)	(0.54)	(0.53)	(0.53)	(0.53)	(0.53)	(0.53)	(0.53)	(0.53)	(0.53)	(0.53)	(0.53)

[a] Risk allele frequency
Population Prevalence of Disease (K) = 0.10
Number of markers on genome-wide platform (M) = 500,000
Overall type I error (α) = 0.05

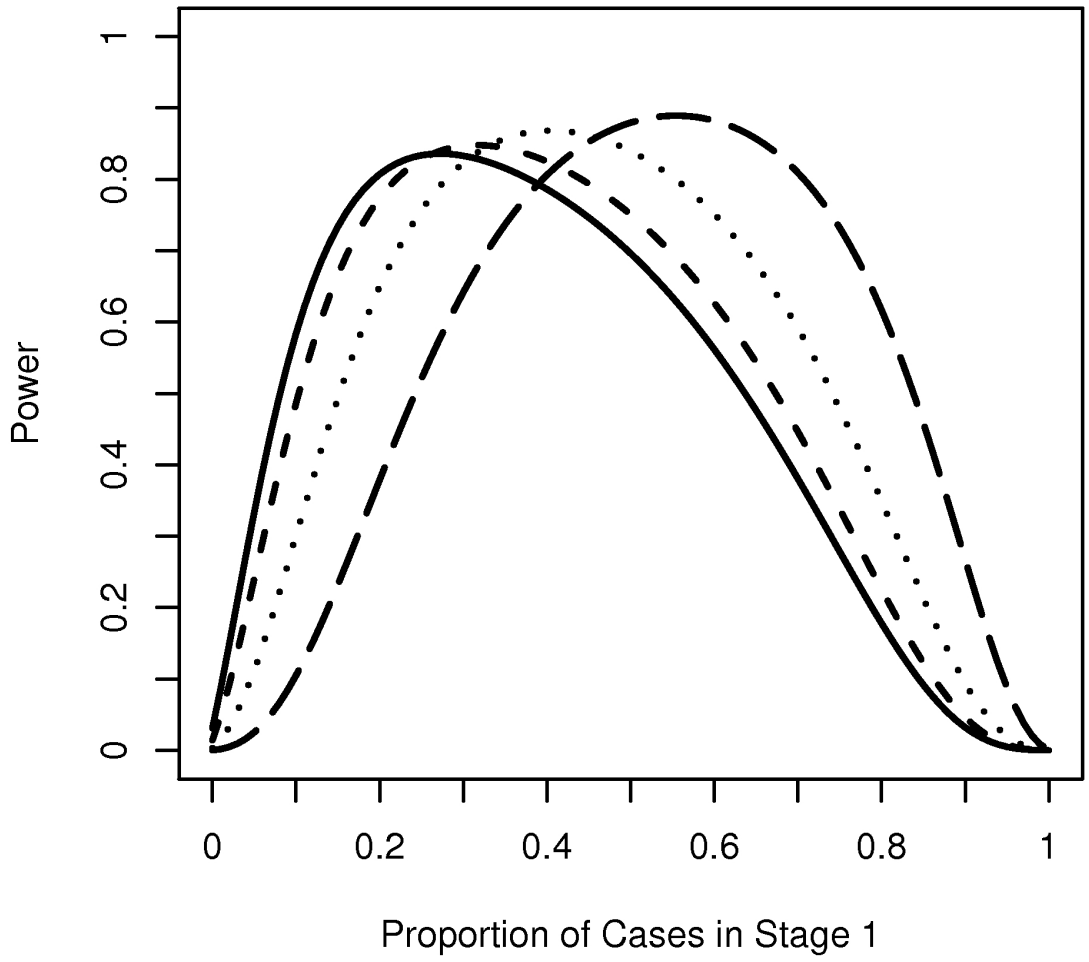

FIGURE 2.1: Power for the Trend Test in 2-Stage Replication-Based GWA Study Designs with 500,000 SNPs Across a Range of Follow-up Platforms, Using 2,000 Cases, 5,000 Public Controls (Stage 1), 2,000 Screened Controls (Stage 2) and Assuming a Multiplicative Model. The different line types reflect the power curves for different follow-up platforms across the possible range of proportion of cases genotyped in stage 1. The follow-up platforms are defined by the number of markers genotyped in stage 2: a) solid line 16,500 SNPs; b) short-dash line 7,500 SNPs; c) dotted line 1,500 SNPs; d) long-dash line 100 SNPs. We assumed the population prevalence of disease (K), the risk allele frequency (f_D), and genetic relative risk (GRR) was 0.10, 0.3, and 1.3, respectively. The maximum power and the corresponding proportion of cases genotyped in stage 1 (at which maximum power occurred) for the various study designs were: a) 16,500 0.836 and 0.27; b) 7,500 0.848 and 0.31; c) 1,500 0.868 and 0.40; d) 100 0.889 and 0.55.

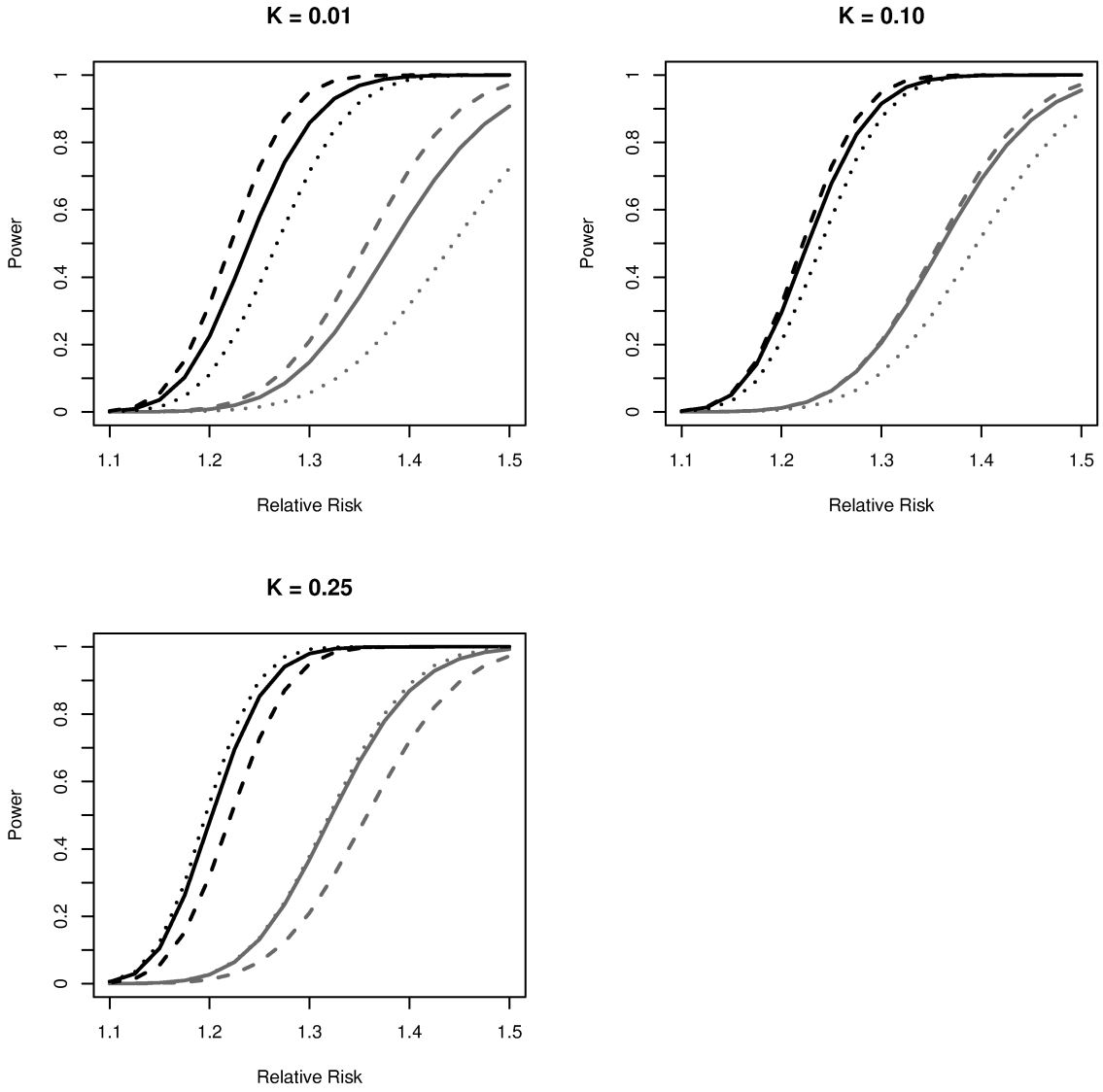

FIGURE 2.2: Power for the Trend Test in 1- and 2-Stage GWA Study Designs Assuming 500,000 Markers, 2,000 Cases, 5,000 Public Controls, and 2,000 Screened Controls. Results are Presented Across a Range of Genotype Relative Risks and Assuming a Multiplicative Risk Model, Risk Allele Frequency (f_D) of 0.1 and 0.5, and Disease Prevalences (K) of 0.01, 0.10, and 0.25. Each panel presents power curves for disease prevalences (K) of 0.01, 0.10, and 0.25. Grey and black lines depict power when the frequency of the disease susceptibility allele (f_D) is 0.1 and 0.5, respectively. Solid lines correspond to the optimal two-stage GWA study based on 5,000 public controls in stage 1 and 2,000 screened controls in stage 2. Dashed lines represent a one-stage GWA study using 5,000 public controls. Dotted lines represent a one-stage GWA study with 2,000 screened controls. Dot-dash lines represent a one-stage GWA study combining 2,000 screened controls with 5,000 public controls. The overall type I error (α) was set at 0.05.

CHAPTER 3

HAPLOTYPE SHARING METHODS IN ASSOCIATION STUDIES

3.1 Introduction

Genome-wide association (GWA) studies offer a promising approach to discover common genetic determinants of disease. Publicly accessible data on human genetic variation from the International HapMap Project (International HapMap Consortium, 2005), plunging genotyping costs, and the availability of high-throughput commercial genome-wide platforms have contributed to its widespread use. GWA studies represent a hypothesis-generating approach since the genomic location of disease susceptibility variants is not assumed, but rather the aim is to uncover these variants (Borecki and Suarez, 2001; Hirschhorn and Daly, 2005). To date numerous common genetic variants have been identified to be associated with common diseases such as type 2 diabetes, prostate cancer and psoriasis.

Barrett and Cardon (2006) evaluated genomic coverage for common and rare SNPs

using HapMap's Phase II and ENCODE (International HapMap Consortium, 2005) data, respectively, in several of Illumina's and Affymetrix's platforms. These platforms were designed to capture common variation. In particular, the Illumina HumanHap300 and Affymetrix 500K panels captured 75% and 65% of the common SNPs, respectively, in Americans of European ancestry (CEU). On the other hand, none of the genome-wide products captured rare SNPs well, at a frequency of less than 10% in the CEU, Yoruba from Nigeria, Japanese from Tokyo, and Han Chinese from Beijing. This deficit in coverage of rare SNPs is still observed as the number of SNPs on commercial genotyping platforms continues to expand. A major limitation for the study of diseases associated with rare variants is that commercial genotyping platforms (particularly Illumina) select SNPs for inclusion on their panels based on available genotype and linkage disequilibrium information in the HapMap Phase I and II samples. Unfortunately, most rare SNPs were missed in HapMap samples because SNP discovery has been limited to a small number of subjects. The 1000 Genomes Project, which significantly expands the number of samples with genomic sequencing information (www.1000genomes.org), is an ongoing project designed to specifically discover rare genetic variants. This discovery should lead to inclusion of many new rare variants on next generation genotyping platforms. Currently there is a renewed interest in identifying rarer functional variants that are associated with disease and these expanded genotyping platforms should facilitate these studies in an economically sensible way. However, disease-specific highly-penetrant-but-rare-founder mutations will likely not be detected during the 1000 Genomes Project sequencing efforts and hence not included on future genotyping platforms.

Due to the computational burden of analyzing large datasets generated from GWA studies using commercial platforms, single SNP tests are the analytical tool of choice (Balding, 2006). However, it has been suggested that haplotype 'blocks' define the se-

quence variation throughout the genome, in which the blocks are more conserved than in other regions (Daly et al., 2001; Jeffreys et al., 2001; Patil et al., 2001; Gabriel et al., 2002). In many circumstances, haplotypes better capture an underlying untyped causal variant than any single genotyped genetic marker. An alternative analytic approach to capture unmeasured genetic risk variants is genotype imputation; however, this approach too relies on the directly causal variants being genotyped in large data sets such as the HapMap samples. Genotype imputation of rare variants, utilizing data available from the 1000 Genomes Project, should increase efficiency/power when evaluating many rare functional variants associated with common disease but such an approach is not likely to be useful for capturing the rare-high-penetrant disease specific founder mutations that likely exist for many rarer disorders. Haplotype analyses are still the best analytic method for detecting founder-mutation disease associations. Ultimately direct high-throughput sequencing may be the solution, particularly if the disease gene location can be significantly narrowed using additional genetic information such as linkage analysis, but this technology is relatively new and is currently prone to high error rates and likely substantial noise (many mutations will be observed) masking any true signals.

Haplotype-based association studies should be particularly useful for identifying susceptibility genes, where susceptibility is conferred by a small number of very rare but highly penetrant variants or mutations that are passed down from generation to generation. These variants/mutations will likely not be identified by the ongoing efforts of the 1000 Genomes Project or other large-scale sequencing efforts that are not specifically focused on the gene of interest. Such rare variants/mutations will typically be relatively recent, as random drift typically influences the frequency of rare mutations to extinction. As a consequence, the haplotype surrounding a rare variant/mutation will be highly conserved due to the relatively small number of generations of recombination

since the variant/mutation occurred. Utilizing extended haplotype sharing between affected individuals around the disease susceptibility variants/mutations has proven to be very powerful for fine-mapping the underlying causal gene for many diseases including ataxia-telangiectasia, Huntington's disease, cystic fibrosis and breast cancer. In all cases, an excess of long-range shared haplotypes between affected individuals, and the break-points in these extended shared haplotypes, pinpointed the location of the susceptibility gene. Success was enjoyed despite the fact that there were many different founder mutations associated with these diseases (allelic heterogeneity), a scenario which can have detrimental effects on power for single SNP studies. One strategy that has previously been shown to be effective is to restrict association studies looking for founder effects to population isolates, where the expected number of unique founder mutations is expected to be considerably less than outbred/mixed populations such as the U.S.. As long-range sequencing becomes more economically feasible and the quality improves, the identification of long-range shared haplotypes between affected subjects should aid in targeting specific regions for sequencing.

There are caveats to haplotype-based approaches. One of which is that in most cases haplotypes are not directly measured and must be statistically inferred using unphased genotype data. Algorithms for population based haplotype inference have been proposed by Excoffier and Slatkin (1995), Clark et al. (1998), and Stephens et al. (2001), among others. Alternatively, in order to empirically determine haplotype phase in the laboratory, molecular haplotyping methods have been developed (McDonald et al., 2002; Odeberg et al., 2002; Burgtorf et al., 2003) though are not widely used because they are not high-throughput, are costly to implement, and have unresolved technical issues (Niu, 2004). For this investigation, haplotype phase was assumed.

The haplotype χ^2 test of association (Sham, 1998) is a traditional approach, in which haplotypes are categorized together if their ordered set of contiguous alleles match at

each genetic marker, the size and location of the window spanning the haplotypes are designated *a priori* by the analyst. For cases and controls, a $2 \times c$ (where c is the unique number of haplotypes) contingency table is then constructed such that disease status and haplotype signatures define the rows and columns, respectively, and a test of no association is conducted. Other traditional haplotype association methods such as that of Clayton (1999) (for family-based studies) and that described by Schaid (2004a) (for case-control studies) are likelihood-based and, as with the haplotype χ^2 test, also require a predefined set of markers for analysis, which can present several problems (Lange and Boehnke, 2004). Assigning windows with too few markers can limit the ability of haplotypes to capture the important variability in the region of interest. Assigning windows that contain too many markers could result in haplotypes with low frequencies (i.e. sparse data) and unnecessarily increase the degrees of freedom. In addition, including too many markers can separate haplotypes containing identical-by-descent segments immediately surrounding the susceptibility variant. All of these limitations can attenuate associations with disease, reducing the power to detect associations if they exist.

As an alternative to the fixed window approach, Van der Meulen and te Meerman (1997a, 1997b) proposed the Haplotype Sharing Statistic (HSS). At the time of their proposal, they motivated their approach with the population genetic assumption that a few of the affected individuals' haplotypes from a founder population not only bear the predisposing disease variant, but also surrounding this locus many alleles are identical by descent (IBD). Nolte and te Meerman (2002) later showed that in comparing two haplotypes at a locus, as the number of identical by state (IBS) marker alleles increases, the probability that the haplotypes are IBD increases. In other words, increased sharing between two haplotypes at a locus suggests that they are IBD. The HSS was designed for nuclear families with one or more affected offspring. A reference

marker is chosen and sharing is assessed upstream and downstream of this location. All possible pairings of haplotypes are considered and for a given pair of haplotypes, the distance recorded represents the length of contiguous matching between alleles at each marker locus surrounding the reference marker. At a given marker locus, alleles match if they are IBS. The HSS is then computed to be the standard deviation among the entire sample of recorded shared distances. Unlike traditional haplotype methods that restrict analysis to a small subset of markers, Van der Meulen and te Meerman offer a data driven approach that allows the use of all available marker data (Lange and Boehnke, 2004). Unlike traditional haplotype-based approaches, the inclusion of additional markers should always result in greater power as highly similar yet not completely identical haplotypes still contribute to the detection of a possible association.

Van der Meulen and te Meerman's HSS reflects an approach in which similarity scores are first generated for all possible pairs of haplotypes, then a summary measure (e.g. the HSS) is computed that incorporates these scores, and lastly statistical significance of the summary measure is determined. Others have proposed alternative methods to score haplotype pairs. Tzeng et al. (2003) and defined the 'matching', 'length', and 'counting' measures for case-control studies. Similar to traditional haplotype methods, these measures require a predefined window of markers to analyze. For a given pair of haplotypes and within the prescribed window, the matching measure assigns a score of 1 if all the alleles match and 0 otherwise, the length measure is the length spanned by the longest continuous interval of alleles IBS, and the counting measure is the number of concordant alleles, which does not require matching alleles to be adjoining. One of the drawbacks of these measures is the specification of the window, as is with traditional haplotype tests.

Lange and Boehnke (2004) developed the conserved haplotype sharing statistic (CHSS) and described it in the context of family trios comprising one affected indi-

vidual and his or her parents, the parents may or may not be diseased. For each pair of haplotypes among all pairings, the CHSS is constructed by evaluating alleles to the left and right of a chosen reference marker. To control for possible genotyping errors and marker allele mutations, one marker mismatch on both sides of the reference marker is allowed, at the expense of a user-defined penalty parameter. Rare alleles that match are given more weight than common alleles, by taking the inverse of the estimated allele frequencies when computing the CHSS. Ambiguous phase and missing marker data are accounted for in the scoring algorithm.

Once similarity scores are generated, then a summary statistic may be computed with the intention of assessing its statistical significance. Lange and Boehnke (2004) introduced the Haplotype Runs Test (HRT) statistics that strictly consider scores from transmitted haplotypes (i.e. 'case' haplotypes) in the family trio setting. Also for the trio design with affected offspring, Bourgain et al. (2000) defined the Maximum Identity Length Contrast (MILC), which measures pairs of haplotypes in the same way as Van der Meulen and te Meerman's (1997a, 1997b) method. However, the MILC contrasts the transmitted and non-transmitted samples of haplotypes differently, by subtracting the mean of the scores formed from all possible haplotype pairs of the non-transmitted haplotypes from that of the transmitted haplotypes. Lange and Boehnke (2004) investigated the power of summing only the transmitted haplotypes versus subtracting off the sum of the non-transmitted from the sum of the transmitted haplotypes. They discovered that the former method was much more powerful, postulating that under the alternative hypothesis, for the groups of transmitted and non-transmitted haplotypes, the within group similarity is high while the between group similarity is low. Thus, the sums of scores corresponding to each group would both have reasonably high values and subtracting the non-transmitted from the transmitted scores would obscure this grouping effect.

Neither of the above mentioned methods of summing the scores include scores from discordant pairs of haplotypes, i.e. a transmitted (case) paired with a non-transmitted (control) haplotype, are not incorporated into the summary statistics. Beckmann et al. (2005) defined a statistic that uses all of the available haplotype similarity measures such that each score is weighted. They motivated the weights with the argument that in comparing haplotypes, the corresponding phenotypes, be they continuous traits or measured dichotomously, that deviate the most from the phenotypic mean are the most influential (Elston et al., 2000; Forrest, 2001). For example, in the case-control setting, as the population frequency of disease becomes more rare, pairs of case haplotypes (i.e. haplotypes that deviate most from the norm) are given more weight than pairs of control or discordant haplotypes. Analogously, pairs of control haplotypes weigh more as disease prevalence becomes more common.

In this investigation, we made slight modifications to Lange and Boehnke's (2004) CHSS statistic in that when considering a pair of haplotypes, the alleles at the reference marker must match in order for the CHSS to build up and downstream and we did not allow any mismatches on either side of the reference marker. In addition, we investigated the power of the Length and Count measures. The Length score was based on Van der Meulen and te Meerman's (1997a, 1997b) scoring method in constructing the HSS that measured the shared genetic distance. For matching alleles, the Count score simply counted the number of matching alleles. Furthermore, Lange and Boehnke created an indicator variable from the CHSS values and a predefined threshold value, aiming to distinguish between haplotypes that have extended sharing (thus, more likely to be IBD) and those that share for shorter stretches. Additionally, a haplotype pair or a small number of pairs that are abundantly similar may dominate the test statistic, so using an indicator variable would guard against this. The problems with this approach are interpreting the meaning of this chosen value and knowing *a priori* an adequate

threshold given the data. To address these issues, thresholds are determined from designated percentiles of the ordered array of scores. Thus, the threshold is data driven and more interpretable. Then, we recode each of the similarity measures in two ways. First, we follow the design of the indicator variable as initially proposed by Lange and Boehnke, though we use the percentile based thresholds. Second, we introduce a second score based on thresholds that instead of assigning zeros to scores that are less than the threshold, we divide these scores by the threshold value to construct normalized scores less than one, and scores greater than or equal to the threshold are recoded to one (in essence resulting in a truncated score with a threshold ceiling value).

For each of these similarity measures and recoded variables we employ the summary methods as previously described and then assess statistical significance using permutation tests, similar to Lange and Boehnke (2004). Finally, we introduce a novel approach that exploits the observation that similar haplotypes form clusters. As opposed to permutation tests that are computationally burdensome, this novel approach is quick and efficient in that contingency tables are constructed using the percentile based thresholds and a p-value is computed with Pearson's χ^2 statistic.

We found that the $\log_{10}$ version of the CHSS outperformed the other reference marker scores, dichotomizing the haplotype sharing scores with a threshold based on percentiles increased power, using fixed windows was detrimental to power, removing rare SNPs and SNPs in high LD with each other was not recommendable, and our novel clustering algorithm had competitive power and was significantly faster than permutation testing, which is desirable for genome-wide scans.

3.2 Methods

We assumed that our sample consisted of an equal number of N cases and N controls, unrelated and independent, for a total of $2N$ subjects and $4N$ haplotypes. All subjects were assumed to be genotyped at M genetic polymorphic markers (for example, the markers could be single nucleotide polymorphisms [SNPs], microsatellites, short tandem repeats, etc.) with no missing data and the markers were ordered by physical location. For the pool of $4N$ total haplotypes, we considered all possible $\binom{4N}{2}$ pairings of haplotypes, such that case (control) haplotypes were paired with other case (control) haplotypes as well as case haplotypes paired with control haplotypes. We assumed Hardy-Weinberg Equilibrium and thus haplotypes within individuals were regarded as independent and also included as a possible pairing. We generated a multitude of sharing statistics for each pair of haplotypes, so that for each sharing statistic there were $\binom{4N}{2}$ scores.

The sharing statistics were constructed based on either a reference marker or fixed window of markers. In the reference marker approach, an initial starting (or "reference") marker was chosen and scores were computed up- and downstream of the reference marker up to, but not including, the first mismatched pair of alleles. Thus, for long stretches of consecutive haplotype sharing, the reference marker approach did not restrict the magnitude of the sharing statistic to a predetermined number of markers. On the other hand, with the fixed window method we specified a region of markers to be considered upon calculating the particular sharing statistic, such that markers outside of this region were not considered.

3.2.1 Reference Marker Approach

We calculated the conserved haplotype sharing statistic (CHSS) as proposed by Lange and Boehnke (2004). Specifically, for a given pair of haplotypes, at the chosen reference marker, r, within the ordered set of markers, $1 \leq r \leq M$, the observed alleles were required to be identical by state (IBS) in order for the CHSS to be constructed, otherwise the CHSS was set to 1. Let A_i^j be the specific allele at marker i, $1 \leq i \leq M$, on haplotype j ($j = 1, 2$) of the haplotype pair and let $\hat{f}_i(A_i^j)$ be the estimated population-based frequency of allele A_i^j computed using the entire sample of case and control alleles. Define markers a and x ($1 \leq a \leq r \leq x \leq M$) such that a (x) is the first marker to the left (right) of the reference marker that fails to match alleles IBS between the two haplotypes. Given that the alleles matched at r, the CHSS was then defined to be the product of the reciprocal of the $\hat{f}_i$'s across the contiguous interval of alleles IBS for a given pair of haplotypes. Namely,

$$\text{CHSS}^r = \prod_{i=a+1}^{x-1} \left(\hat{f}_i(A_i^1) \right)^{-1} \tag{3.1}$$

Since the values of the allele frequencies were between 0 and 1, computing the reciprocal of the $\hat{f}_i$'s gave much greater weight to matching alleles that were more rare (i.e. as $\hat{f}_i \to 0$, $\hat{f}_i^{-1} \to \infty$). In contrast, identical alleles that were common (e.g. $0.20 \leq \hat{f}_i < 1$ corresponded to $1 < \hat{f}_i^{-1} \leq 5$) contributed much less to the magnitude of the CHSS.

We also considered the CHSS in log base 10 space upon motivating the summary statistics of the haplotype scoring measures. The CHSS in (3.1) was then

$$\log_{10}(\text{CHSS}^r) = \sum_{i=a+1}^{x-1} \log\left(\hat{f}_i(A_i^1) \right)^{-1} \tag{3.2}$$

Starting at the reference marker r and for each pair of haplotypes, we calculated the total length of the continuous region over which all markers were IBS. This was similar to the scoring measure used in the Maximum Identity Length Constrast (MILC) statistic as described by Bourgain et al. (2000). In particular,

$$\text{Length}^r = \text{Pos}(x-1) - \text{Pos}(a+1) \tag{3.3}$$

where Pos was the relative position on the chromosome at either marker $x-1$ or $a+1$. For example, Pos could be the physical position in basepairs.

The final reference marker based sharing measure we examined was the Count. As the name implies, for a given pair of haplotypes we counted the number of identical alleles, beginning at r and then moving to the left and right of r until we reached mismatching alleles.

$$\text{Count}^r = \sum_{i=a+1}^{x-1} 1 \tag{3.4}$$

3.2.2 Fixed Window Approach

The haplotype sharing scores based on reference markers described in Section 3.2.1 were also implemented utilizing fixed windows. In other words, a region of markers was defined and analogous measures to the CHSS, Length, and Count were computed within this specified region, in addition to a binary score detailed by Tzeng et al. (2003). For the CHSS and Count we simply considered all matching alleles within this designated segment, regardless of any potential mismatches between markers. The Length measure was defined to be the length of the longest continual interval within the assigned window.

Define the lower and upper boundaries of the fixed window to be w_1 and w_2, respectively, such that amongst the ordered set of M markers, $1 \leq w_1 \leq w_2 \leq M$. The

CHSS under the fixed window approach, analogous to the CHSS based on a reference marker (3.1), was then

$$\text{CHSS}_{w_1}^{w_2} = \prod_{i=w_1}^{w_2} I_{\hat{f}_i}\left(\hat{f}_i(A_i^1)\right)^{-1} \tag{3.5}$$

where the indicator variable $I_{\hat{f}_i}$ was set to 1 if at marker i the alleles for haplotypes 1 and 2 matched (i.e. $A_i^1 = A_i^2$), otherwise (i.e. $A_i^1 \neq A_i^2$) we set $I_{\hat{f}_i}$ to $\hat{f}_i$, which effectively did not increase the size of the CHSS score for alleles that were not IBS. Lange and Boehnke (2004) did not investigate the CHSS using windows, rather the sharing scores included in their report were restricted to scores constructed about a reference marker.

The $\log_{10}$ equivalent of $\text{CHSS}_{w_1}^{w_2}$ was

$$\log_{10}(\text{CHSS}_{w_1}^{w_2}) = \sum_{i=w_1}^{w_2} \log_{10}\left[I_{\hat{f}_i}\left(\hat{f}_i(A_i^1)\right)^{-1}\right] \tag{3.6}$$

The following fixed window scores ($\text{Match}_{w_1}^{w_2}$, $\text{Length}_{w_1}^{w_2}$, and $\text{Count}_{w_1}^{w_2}$) were described by Tzeng et al. (2003).

The Match score for window w_1 to w_2 was

$$\text{Match}_{w_1}^{w_2} = I_{w_1}^{w_2} \tag{3.7}$$

where $I_{w_1}^{w_2}$ was 1 if for a given pair of haplotypes all of the alleles within w_1 and w_2 were identical and 0 if there was at least one discordant pair of alleles (i.e. $A_i^1 \neq A_i^2$).

The window based version of the Length haplotype sharing score given $k = 1, \ldots, c$ continuous segments within the specified window (w_1 to w_2) was

$$\text{Length}_{w_1}^{w_2} = \max_k\left[\text{Pos}(U_k) - \text{Pos}(L_k)\right] \tag{3.8}$$

where U_k and L_k were the upper and lower markers that bounded the k-th continuous interval. We found the largest such interval within w_1 and w_2 and set this to $\text{Length}_{w_1}^{w_2}$

for a given pair of haplotypes.

The Count measure using windows was

$$\text{Count}_{w_1}^{w_2} = \sum_{i=w_1}^{w_2} I_i \tag{3.9}$$

where I_i was an indicator variable defined to be 1 for matching alleles ($A_i^1 = A_i^2$) and 0 for non-identical alleles ($A_i^1 \neq A_i^2$). The $\text{Count}_{w_1}^{w_2}$ counted all alleles IBS within w_1 and w_2 and did not require that matching alleles be adjacent to one another.

3.2.3 Threshold Scores

We constructed threshold scores for each of the haplotype sharing measures based on reference markers (Section 3.2.1) and windows (Section 3.2.2) in order to distinguish between groups of haplotypes, with the exception of the $\text{Match}_{w_1}^{w_2}$ (3.7) which by definition was binary. The motivation was that across the $\binom{4N}{2}$ haplotype pairings there would be a varying degree of sharing scores, where smaller scores represented haplotypes that failed to match or matched for only a few common alleles and larger scores identified haplotypes that matched over an extended set of markers. The threshold scores were aimed at separating these two contrasting sets of shared haplotypes. More importantly, upon summing the haplotype sharing scores (described in Section 3.2.4), a single or small number of haplotype pairs that exhibit a high degree of sharing would not dominate the test statistic under a threshold score.

Lange and Boehnke (2004) proposed applying thresholds on their CHSS. They considered two threshold values, $t = 100$ and $t = 10,000$, for which the modest threshold value of $t = 100$ focused on excess sharing of short or common haplotypes whereas the high threshold of $t = 10,000$ focused on excess sharing of rare or extended haplotypes. We further developed the idea of thresholds by defining thresholds based on percentiles

of the set of $\binom{4N}{2}$ haplotype sharing scores. The use of percentiles allowed the threshold values to be driven by the data, instead of arbitrarily selecting these threshold values, which could be adequate for some data sets but not for others. Furthermore, in contrast to the investigation of Lange and Boehnke (2004) in which the thresholds were employed only for the CHSS, the use of percentiles permitted us to utilize thresholds across a range of haplotype sharing methods and compare their performance in terms of power.

Specifically, to ascertain the threshold values, T_{P_k}, given a set of percentiles (P_k for $k = 1, \ldots, t$) and a haplotype sharing score type (e.g. $\log_{10}$(CHSS), Length, and Count), we ordered the entire set of $\binom{4N}{2}$ scores, $S_{(\ell)}$ for $\ell = 1, \ldots, \binom{4N}{2}$, and determined the location L that demarcated P_k percent of the ordered scores, from (1) to (L). We then defined T_k to be $S_{(L)}$. Since $\log_{10}$ is a one-to-one transformation, in building the threshold based scores, we opted to compute threshold scores for $\log_{10}$(CHSS) and not for CHSS (unlogged).

We defined two types of threshold scores that we used for both the reference marker and window based haplotype sharing measures. The first was binary in nature,

$$I_{\{S_{(\ell)} \geq T_{P_k}\}} = \begin{cases} 1, \, S_{(\ell)} \geq T_{P_k} \\ 0, \, S_{(\ell)} < T_{P_k} \end{cases} \tag{3.10}$$

where $S_{(\ell)}$ could either be a score computed using a reference marker (e.g. $\log_{10}(\text{CHSS}^r)$, Length^r, Count^r) or a fixed window (e.g. $\log_{10}(\text{CHSS}_{w_1}^{w_2})$, $\text{Length}_{w_1}^{w_2}$, $\text{Count}_{w_1}^{w_2}$). We note that the actual value of T_{P_k} depended on the score type under assessment.

The second threshold score weighted the more similar haplotypes (i.e. the pairs of haplotypes that had scores exceeding the corresponding threshold) equally and rendered

the less similar haplotypes on a continuous scale. Specifically,

$$R_{\{S_{(\ell)} \geq T_{P_k}\}} = \begin{cases} 1, \, S_{(\ell)} \geq T_{P_k} \\ \dfrac{S_{(\ell)}}{T_{P_k}}, \, S_{(\ell)} < T_{P_k} \end{cases} \tag{3.11}$$

where $0 \leq \frac{S_{(\ell)}}{T_{P_k}} < 1$. Haplotypes with a high degree of sharing, as determined by the threshold T_{P_k}, were weighted in the same way as the binary threshold score $I_{\{S_{(\ell)} \geq T_{P_k}\}}$. However, for haplotypes that did not meet this threshold criterion, we assigned the ratio $S_{(\ell)}$ to T_{P_k}, the particular score to the threshold. Haplotypes that were fairly similar but did not quite surpass T_{P_k} had a ratio close to 1 while on the contrary haplotypes that did not have any alleles IBS or had a few common alleles had a ratio of 0 or close to 0. We intended that $R_{\{S_{(\ell)} \geq T_{P_k}\}}$ would utilize available information about dissimilar haplotypes, as compared to simply assigning them a 0 which was the approach of $I_{\{S_{(\ell)} \geq T_{P_k}\}}$.

3.2.4 Summary Statistics

We considered three methods of summing the haplotype similarity measures. First, we defined a summary statistic that compared the haplotype sharing amongst the cases with that of the controls. Nolte et al. (2007) investigated such a statistic for which they summed the pairwise haplotype sharing in the cases and controls separately and then calculated the difference in these sums in the construction of their test statistic, the haplotype-sharing statistic (HSS). This was in contrast to their earlier proposed HSS (Van der Meulen and te Meerman, 1997a, 1997b) that utilized all pairs of "case" haplotypes (Van der Meulen and te Meerman based their report on family data such that "case" haplotypes were haplotypes transmitted from the parents to the offspring). Bourgain et al. (2000) also applied this approach of subtracting out the effect of the

controls (i.e. non-transmitted haplotypes) from the cases (i.e. transmitted haplotypes). Nolte et al. (2007) claimed that such an approach, as applied to their HSS, could correct for linkage disequilibrium (LD) other than that caused by the disease mutation. Specifically, we defined this sum in (3.14) as

$$\text{Sum}^{\text{Case}} = \sum_{i=1}^{4N-1} \sum_{j=i+1}^{4N} I_{\text{Case},i} I_{\text{Case},j} S_{ij} \tag{3.12}$$

$$\text{Sum}^{\text{Con}} = \sum_{i=1}^{4N-1} \sum_{j=i+1}^{4N} I_{\text{Con},i} I_{\text{Con},j} S_{ij} \tag{3.13}$$

$$\text{Sum}^{\text{Diff}} = \text{Sum}^{\text{Case}} - \text{Sum}^{\text{Con}} \tag{3.14}$$

where Sum^{Case} and Sum^{Con} were the sum of the given haplotype sharing scores S_{ij} (as described in Sections 3.2.1 and 3.2.2), i and j indexed the haplotypes of the pairing, $I_{\text{Case},i(j)}$ was set to 1 if the $i(j)$-th haplotype was from a case subject and 0 if it was from a control, and likewise $I_{\text{Con},i(j)}$ was the indicator variable for the control haplotypes. Assuming Hardy-Weinberg equilibrium, case and control haplotypes within subjects were included in Sum^{Case} and Sum^{Con}.

Lange and Boehnke (2004) investigated the power of Sum^{Case} and Sum^{Diff} in the context of the parent-parent-affected offspring trio design in which the "case" and "control" haplotypes were those haplotypes transmitted and not transmitted from the parents to the offspring, respectively. They discovered that Sum^{Case} was much more powerful, postulating that under the alternative hypothesis, for the groups of transmitted and non-transmitted haplotypes, the within group similarity was high while the between group similarity was low. Thus, Sum^{Case} and Sum^{Con} would both have reasonably high values and computing their difference (Sum^{Diff}) would obscure this grouping effect. For the second summary measure, we extensively studied the power of Sum^{Case} in order to examine closely Lange and Boehnke's claim.

Lastly, the third summary measure we considered was the general approach of Mantel's statistics for space-time clustering (Mantel, 1967) to correlate genetic and phenotypic similarity. Beckmann et al. (2005) proposed the use of Mantel's statistic whilst defining the phenotypic similarity. In contrast to Sum^{Case} and Sum^{Diff} that do not utilize all $\binom{4N}{2}$ scores, the Mantel statistic applies a weighting measure across all scores. Namely,

$$M = \sum_{i=1}^{4N-1} \sum_{j=i+1}^{4N} S_{ij} Y_{\text{Subj}_i, \text{Subj}_j} \tag{3.15}$$

where $Y_{\text{Subj}_i, \text{Subj}_j}$ was the phenotypic similarity for two haplotype copies i and j in which the phenotypes were derived from subjects Subj_i and Subj_j. Specifically,

$$Y_{\text{Subj}_i, \text{Subj}_j} = (y_{\text{Subj}_i} - \mu)(y_{\text{Subj}_j} - \mu) \tag{3.16}$$

$Y_{\text{Subj}_i, \text{Subj}_j}$ was the mean corrected product where μ denoted the mean of the phenotype and y_{Subj_i} and y_{Subj_j} the phenotypes of subjects Subj_i and Subj_j. The motivation behind this definition was that either similar or different haplotype pairs farthest from the mean μ would be the most influential (Elston et al., 2000; Forrest, 2001). There were several possibilities to define μ and in the scenario of a binary phenotype such as case/control status in our unrelated cases and controls design, we set μ to be the disease prevalence and $y_{\text{Subj}_{i(j)}}$ was 1 if $\text{Subj}_{i(j)}$ was a case and 0 if a control. μ was a parameter that weighed the three possible comparisons between pairs of haplotypes, i.e. 1) both were from affected/exposed individuals; 2) both were from control individuals; 3) the haplotype pair was discordant, i.e. one was from an affected/exposed individual and the other came from a control individual. For rare diseases, disease prevalence is

close to 0 (i.e. $\mu \approx 0$), and it follows that

$$Y_{\text{Subj}_i, \text{Subj}_j} = (y_{\text{Subj}_i} - \mu)(y_{\text{Subj}_j} - \mu) = \begin{cases} \approx 1 & \text{if both haplotypes } i \text{ and } j \text{ were from affected individuals} \\ \approx 0 & \text{if both haplotypes } i \text{ and } j \text{ were from unaffected individuals} \\ < 0 & \text{if the haplotypes } i \text{ and } j \text{ were from discordant individuals} \end{cases}$$

On the other hand, the more common the disease, i.e. as disease prevalence approaches 50% ($\mu \rightarrow 0.5$), the values of $Y_{\text{Subj}_i, \text{Subj}_j}$ for concordant affected and unaffected haplotype pairs converge from 1 and 0, respectively, to 0.25, whereas scores for discordant haplotype pairs (i.e. one haplotype comes from an affected individual and the other from an unaffected) become more negative and reach -0.25. In this scenario, concordant pairs, regardless of affection status, are scored equally and contrasted to discordant pairs, thereby testing whether they have a tendency to share protective haplotypes (Beckmann et al., 2005).

3.2.5 Significance Estimation of the Summary Statistics via Permutation Testing

The asymptotic distributions of the various haplotype sharing scores generated by the reference marker approach (Section 3.2.1) were mathematically intractable due to the nature of these scores in which sharing was assessed for the length of the haplotype pairs until a mismatch was encountered. Furthermore, we could not satisfactorily approximate these distributions with any known distribution(s). Consequently, we empirically estimated statistical significance of the summary statistics described in Section 3.2.4 via permutation testing, for the sets of summary statistics pertaining to

both reference marker and fixed window scores (Section 3.2.2).

In particular, we randomly shuffled the affection status labels across all N cases and N controls. To save computational time and resources, for each score type we did not regenerate the $\binom{4N}{2}$ haplotype pairing scores, but rather we simply reassigned each score with its corresponding pair of labels resulting from the given random shuffle. We then computed the summary statistics, Sum^{Case}, Sum^{Diff}, and M (Equations 3.12, 3.14, and 3.15, respectively) according to the permuted affection status labels. We carried this out B times, resulting in a set of $t^{(1)}, t^{(2)}, \ldots, t^{(B)}$ permutation summary statistics for each summary statistic type relating to a given score. We defined the permutation p-value to be

$$p = \frac{\sum_{b=1}^{B} I(t^{(b)} \geq t)}{B} \tag{3.17}$$

where t was the observed summary statistic and I was the indicator function. For $t^{(b)} = t$, with one-half probability we set I to be one and zero otherwise.

3.2.6 Single Marker χ^2 Test of Association

We compared the performance of the haplotype sharing measures and summary statistics with the single marker χ^2 test of association. At a given reference marker, r, we performed an allelic association test of 1 degree of freedom. We constructed a 2×2 contingency table in the following manner. Disease status constituted the rows and the alleles at r (e.g. d and D) were classified in the columns, such that each subject contributed two counts to any given cell. Specifically, a homozygote dd was counted twice in the d allele column, similarly for a homozygote DD in the D allele column, and a heterozygote Dd contributed one count to the d column and another to the D column. We employed Pearson's 1 degree of freedom χ^2 statistic, Q_P, to test the null hypothesis of no association. We assumed Hardy-Weinberg equilibrium (HWE) in both cases and controls.

3.2.7 Haplotype χ^2 Test of Association

We also assessed the power of the haplotype χ^2 test of association. We used the same fixed windows that we defined for the fixed window scores (Section 3.2.2) for the haplotype χ^2 test. In particular, given a fixed window with lower and upper boundaries of w_1 and w_2, we searched the resulting pool of haplotypes for distinct haplotype signatures and each signature was assigned a row in a $R \times 2$ contingency table, where R was the total number of rows (i.e. unique haplotype patterns) and there were 2 columns that classified affection status. All haplotypes were then categorized into their corresponding haplotype row and affection status column (i.e. haplotypes arising either from affected or unaffected individuals were labeled as such). The cells of the $R \times 2$ table contained the counts of the categorized haplotypes.

To test the null hypothesis of no association, we computed the χ^2 statistic that had approximately a χ^2 distribution with $R-1$ degrees of freedom. Beforehand we checked for small haplotype row totals so that the number of rows (and hence the degrees of freedom) would not be excessively large if many rare haplotypes were present in the sample. Furthermore, the χ^2 approximation was more appropriate when expected cell sizes were 5 or greater. We implemented 2 approaches when haplotype row totals were sufficiently small, i.e. < 10. First, we removed these rows entirely from the table ("Delete" method). Second, we pooled together all rows with small sample sizes ("Collapse" method).

3.2.8 An Alternative to Permutation Testing: a Quick and Efficient Clustering Algorithm for Significance Estimation of Haplotype Sharing Measures

We developed methods to assess statistical significance of excess haplotype sharing amongst the cases as compared to the controls. Aside from the biological motivation, these methods were computationally quick and efficient.

We investigated the haplotype sharing amongst 60 unrelated individuals (120 haplotypes) of the CEPH sample (Utah Residents with Northern and Western European Ancestry), genotyped as part of the HapMap project. Across all $\binom{120}{2} = 7,140$ pairs of haplotypes, we examined the number of contiguous alleles starting at a reference marker in the PHB gene region. We chose 2 markers, one that was relatively common (rs2233667, estimated minor allele frequency [MAF] = 0.328) and the other that was rare (rs882031, estimated MAF = 0.025). Surprisingly, a fair amount of unrelated independent haplotypes shared a considerable number of adjacent alleles, regardless of the MAF of the initial reference marker. We therefore hypothesized that case haplotypes as well as control haplotypes in themselves shared a unique set of haplotype patterns. This was similar to Lange and Boehnke's 2004 assertion that under the alternative hypothesis, for the groups of transmitted and non-transmitted haplotypes in a family-based study design, the within group similarity is high while the between group similarity is low. Thus, we were prompted to design an approach that allowed affected and unaffected haplotypes to cluster according to a relative measure of similarity and to formally test the statistical significance of the observed clustering in the cases versus the controls.

We carried out the clustering algorithm in the following manner. Begin cluster formation with an arbitrary single haplotype, $h_{c_1,1}$, where c_1 notated the first cluster

being formed. The thresholds, T_{P_k}'s, specific for a haplotype sharing score type (e.g. reference marker and fixed window based $\log_{10}$(CHSS), Length, and Count) described in Section 3.2.3 served as the cutoffs for which we designated haplotypes to be members of a particular haplotype grouping.

1. Amongst the haplotype sharing scores pertaining to the haplotype pairings with the initial haplotype in the cluster, search for any haplotype that meets or exceeds T_{P_k} and include them in the cluster. Note, for $h_{c_1,1}$ the threshold scores searched is a subset of $4N-1$ scores out of a total of $\binom{4N}{2}$ haplotype pairing scores for $2N+2N=4N$ case and control haplotypes.

2. For each of the haplotypes entered at Step 1, search for any haplotype(s) to further include in the cluster, based on the corresponding subset of haplotype sharing scores.

3. Once the cluster can no longer include additional haplotype members, begin building another cluster starting with an arbitrary single haplotype, granted that there are haplotypes that have not yet been grouped. Repeat Steps 1 and 2 for new clusters to be formed.

4. As soon as no other clusters can be formed after iterating through Steps 1 through 3 and haplotypes remain that have not been assigned to any of the clusters, place them in an "other" bin to be subsequently assessed.

In Steps 1 and 2, we searched scores corresponding to haplotypes that have not yet been clustered, so haplotypes were not counted more than once. As a result, this also saved computational time and resources when searching.

To illustrate the clustering algorithm, consider the following example. We begin building a cluster with an arbitrary haplotype, $h_{c_1,1}$, for the reference marker based $\log_{10}$(CHSS). We choose P_k to be the 99-th percentile, so the threshold value is T_{99}.

1. 3 other haplotypes ($h_{c_1,2}, h_{c_1,3}, h_{c_1,4}$) have $\log_{10}$(CHSS) values $\geq T_{99}$ from their pairings with $h_{c_1,1}$, so these are included in cluster c_1.

2. (a) 2 other haplotypes ($h_{c_1,5}$ and $h_{c_1,6}$) have $\log_{10}$(CHSS) values $\geq T_{99}$ from their pairings with $h_{c_1,2}$, so are included in cluster c_1.

   (b) 3 other haplotypes ($h_{c_1,7}$, $h_{c_1,8}$, and $h_{c_1,9}$) have $\log_{10}$(CHSS) values $\geq T_{99}$ from their pairings with $h_{c_1,3}$, so are included in cluster c_1.

   (c) No haplotypes have $\log_{10}$(CHSS) values $\geq T_{99}$ from their pairings with $h_{c_1,4}$, so no additional haplotypes are included in cluster c_1.

3. Amongst the $\log_{10}$(CHSS) values computed from the haplotype pairings with each of the haplotypes entered in Step 2 ($h_{c_1,5}, h_{c_1,6}, \ldots, h_{c_1,9}$), no other $\log_{10}$(CHSS) values were $\geq T_{99}$, therefore the construction of cluster c_1 is complete and consists of haplotypes $h_{c_1,1}, h_{c_1,2}, \ldots, h_{c_1,9}$. We begin building another cluster, c_2, starting with arbitrary haplotype $h_{c_2,1}$.

4. After iterating through Steps 1 to 3 two more times, there are 3 clusters, c_1, c_2, and c_3, each containing haplotypes ($h_{c_1,1}, h_{c_1,2}, \ldots, h_{c_1,9}$), ($h_{c_2,1}, h_{c_2,2}, \ldots, h_{c_2,5}$), and ($h_{c_3,1}, h_{c_3,2}, \ldots, h_{c_3,24}$). However, there are still $4N - (9 + 5 + 24 = 38)$ haplotypes remaining that were not clustered, and will be placed in the "other" bin.

We employed various methods to handle clusters containing few haplotypes and haplotypes that were categorized into the "other" bin. We postulated that these rare haplotypes could potentially provide useful information in discerning associations between clusters and affection status, therefore we assessed the performance of incorporating these rare haplotypes compared to removing them entirely from the analysis.

The first technique we implemented was to not attempt to regroup "other" haplotypes into any of the existing clusters ("No Regrouping"). We defined minimum cluster

sizes (size$_i$ for $i = 1, \ldots, N_{\text{sizes}}$) and if the number of haplotypes in a given cluster was not greater than or equal to size$_i$ then we placed all of the haplotypes in this cluster in the "other" bin.

The second method was to attempt to regroup "other" haplotypes into any of the clusters ("Regrouping") in a 3 step process. First, we regrouped "other" haplotypes per criteria which we discuss below. Small clusters that did not fulfill size$_i$ could possibly be expanded at this step. Second, we imposed the cluster size constraints, size$_i$, across all clusters. Third, we attempted to regroup "other" haplotypes once again, since some clusters may have moved to the "other" bin in the previous step.

Our strategy to regroup "other" haplotypes into clusters was the following. For each of the "other" haplotypes, we inspected all of the scores from the pairings with the haplotypes already in clusters and found the maximum score. We regrouped the "other" haplotype into the cluster in which the maximum score resided. We required that the maximum score stemmed from a haplotype that was not originally in the "other" bin. If more than one maximum score was found in multiple clusters, we did not regroup the "other" haplotype in question.

In the third method, we collected all of the small clusters (i.e. all of the clusters that were not as large as size$_i$) into one group instead of recategorizing them into the "other" bin ("Small Cluster Row"). We did not attempt to regroup "other" haplotypes into clusters for this method.

Once the clusters were created by way of the "No Regrouping", "Regrouping", and "Small Cluster Row" methods, we constructed $R \times 2$ contingency tables where the number of rows, R, represented the number of clusters and the 2 columns categorized affection status (i.e. if a haplotype originated from an affected or unaffected individual). We cross classified the clusters by affection status in order to examine if case haplotypes grouped together differently than control haplotypes. Regardless if such a difference

existed or not, the $R \times 2$ tables characterized across all of the clusters the frequency at which the case and control haplotypes congregated based on a quantifiable measure of haplotype sharing.

Similar to how we tested for association in the $R \times 2$ tables of the haplotype χ^2 test (Section 3.2.7), we computed the χ^2 statistic with $R-1$ degrees of freedom. We assessed the performance of including and removing haplotypes that did not assemble into any clusters, which were plausibly the rare haplotypes, by either keeping or removing the "other" group for each of the 3 methods discussed above ("Keep" or "Delete"). For the Delete method, we removed entirely the row of "other" haplotypes, given that such a row existed and that deleting the "other" row did not result in a table with 0 degrees of freedom (i.e. a table with 1 row). On the other hand, for the Keep method, we simply kept in the "other" row when calculating the χ^2 statistic.

Finally, we formed 2×2 tables for which affection status defined the columns and the 2 rows consisted of the aggregated collection of clusters and the group of "other" haplotypes. We did not attempt to regroup "other" haplotypes into the cluster row. We computed the χ^2 statistic to assess statistical significance.

3.2.9 Illumina's iControlDB Public Resource: Acquisition, Cleaning, and Phasing of Genotype Data from Genome-wide Platforms

For the purposes of simulation, described in subsequent sections (Section 3.2.10), we obtained genotypes from a total of 5,444 subjects with diverse ethnic backgrounds (e.g. Caucasians, African-Americans, Hispanics/Latinos, Asians, American Indians), genotyped on Illumina's HumanHap550 Genotyping BeadChip. We downloaded the data with the Illumina iControlDB Client (version 1.1.2.0) upon agreeing to the Illumina

Genotyping Control Database Download Agreement and applying for an Illumina iCom account at www.illumina.com. Control subjects genotyped on versions 1 and 3 of the HumanHap550 platform were available, each comprising 2,990 and 2,454 individuals, respectively.

There were a total of 555,352 and 561,466 genome-wide markers in versions 1 and 3, respectively. Based on the marker names, we found 545,080 markers in common to both versions and 10,272 markers to be in version 1 but not in version 3. Of these discrepant markers, 5,109 SNPs had been renamed since version 1, according to a batch query we performed on dbSNP (www.ncbi.nlm.nih.gov/projects/SNP), and 5,023 of these SNPs in version 1 were in fact in version 3. We recovered these SNPs in common by renaming the older version 1 SNPs with the most current names as they appeared in version 3, which brought the total number of markers in common to both versions 1 and 3 to 550,103. We note that we were unable to recover 16,386 markers that resided in version 3 but not in version 1.

We studied the gene regions BRCA1 and PHB, both on chromosome 17 (we motivate the study of these regions in Section 3.2.10). Therefore, we restricted the genome-wide iControlDB data to chromosome 17. Ten of the SNPs on chromosome 17 (rs2469786, rs1072101, rs2674954, rs11541311, rs2957407, rs3999623, rs4790958 rs1642220, rs2898645, and rs692161) were recorded in version 1 on the opposing complementary strand as version 3, so amongst the subjects in version 1 we recoded these SNPs in accordance with version 3.

We further subsetted the chromosome 17 data to include control subjects who reported to be Caucasian although a small proportion of individuals reported to be of mixed Caucasian ancestry whom we excluded. The Caucasian data was used for subsequent quality control procedures, phasing, and power analyses. In PLINK version 0.99s (Purcell, 2007; Purcell et al., 2007), we carried out the following quality control

procedures. We computed the estimated proportion, $\hat{\pi}$, of alleles shared identity-by-descent (IBD) for all pairwise comparisons of control subjects, in order to locate and remove potentially related individuals such that the sample would consist of independent controls. We also tested for HWE and calculated minor allele frequencies and the rates of missing genotypes by SNP and subject. We then removed individuals and SNPs that were missing more than 5% of their genotypes. SNPs with relatively low MAFs were allowed to remain in the data set. For pairs of subjects who were related to some degree, i.e. $\hat{\pi} \geq 0.2$, we arbitrarily removed one of the related members. Lastly, we made use of PLINK's recoding facilities to output the resulting scrubbed genotype data in fastPHASE format for which we inferred phase and reconstructed haplotypes with fastPHASE version 1.2.3 (Scheet and Stephens, 2006).

3.2.10 Data Simulation to Assess the Power of the Competing Tests

In order to evaluate the performance of the haplotype sharing measures and formal tests, we generated simulated data sets from the pool of reconstructed iControlDB control haplotypes, subsetted on 2 gene regions of particular interest to us, BRCA1 (breast cancer susceptibility gene 1, early onset) and PHB (prohibitin). Both genes are larger than average spanning 81.16 and 10.82 kilobases, respectively, and are on the q arm of chromosome 17 (17q21; BRCA1: base pair [bp] positions 38,449,840 to 38,530,994; PHB: bp positions 44,836,419 to 44,847,241; bp locations based on NCBI B36 assembly / dbSNP b126). BRCA1 has been extensively researched with regard to breast cancer onset in women (National Cancer Institute, 2009a; National Cancer Institute, 2009b). PHB is thought to be a tumor suppressor and involved in sporadic breast cancer (National Center for Biotechnology Information, Entrez Gene, 2009). Aside from the biological relevance in studying BRCA1 and PHB, we were able to

compare and contrast the power of the haplotype sharing measures, summary statistics, clustering algorithm, and traditional approaches when applied to genomic segments of relatively high and low LD as observed in BRCA1 and PHB, respectively.

We simulated phased SNP genotype data across $N = 200$ cases and $N = 200$ controls (unrelated and independent) for a highly penetrant and rare disease by assuming a disease prevalence (K) of 0.001, recessive genetic mode-of-inheritance risk model, genotype relative risk (GRR) of 750, and rare disease allele frequency (f_D) of 0.0125 or 0.02, depending on the pool of haplotypes carrying the disease allele, explained in greater detail below. In order to simulate data sets with these assumed parameters, we sought to explicitly define the genotype probabilities conditional on affection status, i.e. for the cases $\Pr(dd|\text{case})$, $\Pr(Dd|\text{case})$, and $\Pr(DD|\text{case})$, and for the controls $\Pr(dd|\text{control})$, $\Pr(Dd \mid \text{control})$, and $\Pr(DD \mid \text{control})$, where D represented the predisposing disease allele and d the non-causal variant. With these conditional genotype probabilities, we randomly determined the number of cases and controls with a specified genotype (dd, Dd, or DD) and then for each case and control we randomly sampled without replacement entire haplotypes from 2 distinct pools of haplotypes, according to the assigned genotype. One pool consisted of 6,170 haplotypes that were phased and reconstructed from the iControlDB Caucasian control subjects (please refer to Section 3.2.9) and hypothetically carried the non-disease d allele (call it the "d haplotype pool"), whereas the other pool comprised haplotypes hypothetically harboring the disease D allele (call it the "D" haplotype pool). For example, if a case (or a control, for that matter) was randomly assigned a Dd genotype, then we would randomly select without replacement a haplotype from the D and d pools so as to construct the paternal and maternal chromosomal segments. For homozygous cases and controls, i.e. subjects with either dd or DD genotypes, we would randomly sample twice from the d or D haplotype pools, respectively.

For the set of power analyses based on the BRCA1 gene region, the haplotypes in the d and D pools consisted of the SNPs present on the Illumina HumanHap550 marker platform at BRCA1. There were 12 such SNPs beginning at rs8176273 (bp location 38,465,179) and ending at rs799923 (bp location 38,505,457). We selected an additional 151 SNPs up and downstream of BRCA1, for a total of 314 SNPs spanning an approximate 3 megabase region (rs3744786 to rs9891016 corresponding to bp positions 36,876,889 to 39,911,011). There were 6 markers covering PHB ranging from rs1049620 (bp location 44,836,513) to rs2277636 (bp location 44,847,176). Similar to BRCA1, we chose an approximate 3 megabase segment centered about PHB, consisting of a total of 486 SNPs (rs7220419 to rs9905480; bp locations 43,263,404 to 46,291,064) where 240 SNPs were located up and downstream of PHB.

The computational details are as follows. Assuming HWE, f_D, and the non-disease allele frequency ($f_d = 1 - f_D$), we computed genotype probabilities as $f_{dd} = f_d^2$, $f_{Dd} = 2 \cdot f_D \cdot f_d$, and $f_{DD} = f_D^2$. The disease prevalence, K, can be written as

$$K = \Pr(\text{case} \mid dd) \cdot f_{dd} + \Pr(\text{case} \mid Dd) \cdot f_{Dd} + \Pr(\text{case} \mid DD) \cdot f_{DD} \tag{3.18}$$

The GRR under a recessive genetic mode-of-inheritance risk model was defined to be

$$GRR = \frac{\Pr(\text{case} \mid DD)}{\Pr(\text{case} \mid dd)} = \frac{\Pr(\text{case} \mid DD)}{\Pr(\text{case} \mid Dd)} \tag{3.19}$$

It followed from Equation 3.19 that

$$\Pr(\text{case} \mid dd) = \Pr(\text{case} \mid Dd) = \frac{\Pr(\text{case} \mid DD)}{GRR} \tag{3.20}$$

which we substituted into Equation 3.18 to yield

$$\begin{aligned} K &= \frac{\Pr(\text{case} \mid DD)}{GRR} \cdot f_{dd} + \frac{\Pr(\text{case} \mid DD)}{GRR} \cdot f_{Dd} + \Pr(\text{case} \mid DD) \cdot f_{DD} \\ &= \Pr(\text{case} \mid DD) \cdot \left(\frac{f_{dd}}{GRR} + \frac{f_{Dd}}{GRR} + f_{DD} \right) \end{aligned}$$

and rearranging terms gave us the penetrance of DD

$$\text{Pen}(DD) = \Pr(\text{case} \mid DD) = \frac{K}{\frac{f_{dd}}{GRR} + \frac{f_{Dd}}{GRR} + f_{DD}} \tag{3.21}$$

As for the penetrances of dd and Dd, we substituted Equation 3.21 into Equation 3.20

$$\text{Pen}(dd) = \text{Pen}(Dd) = \frac{K}{f_{dd} + f_{Dd} + f_{DD} \cdot GRR} \tag{3.22}$$

where $\text{Pen}(dd)$ and $\text{Pen}(Dd)$ were $\Pr(\text{case}|dd)$ and $\Pr(\text{case}|Dd)$, respectively. Lastly, we computed the genotype probabilities conditional on affection status using the following general relationships for the cases and controls derived from Bayes' theorem

$$\begin{aligned} \Pr(\text{genotype} \mid \text{case}) &= \frac{\Pr(\text{case} \mid \text{genotype}) \cdot f_{\text{genotype}}}{K} \\ \Pr(\text{genotype} \mid \text{control}) &= \frac{\Pr(\text{control} \mid \text{genotype}) \cdot f_{\text{genotype}}}{1 - K} \end{aligned}$$

where "genotype" was dd, Dd, or DD and $\Pr(\text{control} \mid \text{genotype}) = 1 - \Pr(\text{case} \mid \text{genotype})$. Specifically, the probabilities of the cases were

$$\begin{aligned} \Pr(DD \mid \text{case}) &= \frac{f_{DD}}{\frac{f_{dd}+f_{Dd}}{GRR} + f_{DD}} \\ \Pr(Dd \mid \text{case}) &= \frac{f_{Dd}}{f_{dd} + f_{Dd} + GRR \cdot f_{DD}} \\ \Pr(dd \mid \text{case}) &= \frac{f_{dd}}{f_{dd} + f_{Dd} + GRR \cdot f_{DD}} \end{aligned} \tag{3.23}$$

and for the controls were

$$\begin{aligned}
\Pr(DD \mid \text{control}) &= \frac{\left(1 - \frac{K}{\frac{f_{dd}+f_{Dd}}{GRR}+f_{DD}}\right) \cdot f_{DD}}{1-K} \\
\Pr(Dd \mid \text{control}) &= \frac{\left(1 - \frac{K}{f_{dd}+f_{Dd}+GRR \cdot f_{DD}}\right) \cdot f_{Dd}}{1-K} \\
\Pr(Dd \mid \text{control}) &= \frac{\left(1 - \frac{K}{f_{dd}+f_{Dd}+GRR \cdot f_{DD}}\right) \cdot f_{dd}}{1-K}
\end{aligned} \tag{3.24}$$

Therefore, the conditional genotype probabilities could be explicitly computed since we assumed values for K and GRR and we calculated f_{dd}, f_{Dd}, and f_{DD} from the assumed value of f_D.

From the Equations in 3.21 and 3.22, the penetrances of our simulated data under the assumed parameters (recessive genetic mode-of-inheritance risk model, $K = 0.001$, $GRR = 750$, and $f_D = 0.0125$) were $\text{Pen}(dd) = \text{Pen}(Dd) = 0.000895$ and $\text{Pen}(DD) = 0.671$. For the controls, $\Pr(\text{control} \mid dd) = \Pr(\text{control} \mid Dd) = 0.999$ and $\Pr(\text{control} \mid DD) = 0.329$. In other words, if an individual had 2 copies of the predisposing disease allele (DD), there was a firm chance of developing the disease. On the other hand, if an individual did not have 2 copies of the disease allele (dd or Dd), with a very high probability the individual would not contract the disease.

Moreover, our present investigation involved the unrelated case-control study design. Therefore, given our sample of $N = 200$ cases and $N = 200$ controls, their genotype probabilities calculated from the set of Equations in 3.23 and 3.24 were $\Pr(dd|\text{case}) = 0.873$, $\Pr(Dd|\text{case}) = 0.022$, and $\Pr(DD|\text{case}) = 0.105$ amongst the cases and $\Pr(dd \mid \text{control}) = 0.975$, $\Pr(Dd \mid \text{control}) = 0.025$, and $\Pr(DD \mid \text{control}) = 0.000051$ amongst the controls. As an example, one of the realized simulated data sets consisted of the following counts for the cases: 177 (dd), 3 (Dd), and 20 (DD) and for the controls: 197 (dd), 3 (Dd), and 0 (DD).

We further simulated sampling from a founder population in which random mutations in the BRCA1 and PHB gene regions were induced and each haplotype carrying the causal allele was followed throughout a number of generations and recombination events. Specifically, for each simulated data set we generated a founder pool of haplotypes that we designated as the "D haplotype pool" discussed above. We randomly selected 1 or 5 loci within BRCA1 and PHB as the mutation site(s) and for each haplotype we simulated recombination events across 20 or 100 generations under a Poisson process. The recombination simulation was as follows. We randomly sampled a haplotype from the 6,170 haplotypes in the "d haplotype pool" and randomly designated a location for the mutation anywhere within the BRCA1 or PHB gene which did not necessarily have to be at a locus that was genotyped. This mutation was followed after each meiosis. We assumed that crossovers occurred randomly and independently over the entire chromosome (i.e. no interference), which essentially was the Haldane mapping function. In order to determine the number of crossover events for a given meiosis, we multiplied the number of generations (20 or 100) by the recombination fraction, θ, and used this expected number of crossovers as the λ input parameter of the Poisson distribution to draw a random variate. We assessed θ for the designated regions centered about BRCA1 and PHB by first interpolating the sex-averaged map positions in centiMorgans (cM) of the physical positions (in basepairs) that bounded the given gene region via the Rutgers Map Interpolator web application (compgen.rutgers.edu/old/map-interpolator; Matise et al., 2007). Due to the non-additivity of θ, we first calculated the difference of the bounding map positions which gave us the number of cMs for the entire region. We then converted this estimate of additive map distance, x, into the non-additive θ by use of the Haldane map function (Haldane and Smith, 1947),

$$\theta = \tfrac{1}{2}[1 - \exp(-2|x|)]$$

Thus, the single Poisson input parameter (i.e. the expected number of crossovers across 20 or 100 generations) was computed as $\lambda = \theta \cdot 20$ or $\lambda = \theta \cdot 100$, in order to generate a Poisson random variate that designated the number of crossovers for a given meiosis.

Subsequently, we randomly assigned the crossover sites according to the LD map scaled in linkage disequilibrium units (LDUs). The LD maps for BRCA1 and PHB were constructed with the program LDMAP (cedar.genetics.soton.ac.uk/public_html/helpld.html; Maniatis et al., 2002) and plotting the LDUs against the physical map in base pairs revealed a pattern of plateaus and steps. The plateaus signified regions of low haplotype diversity or "LD blocks" whereas the steps reflected recombination hot-spots. We separately read into the LDMAP program the entire set of 312 and 486 SNPs centered about BRCA1 and PHB, respectively, across all 6,170 iControlDB haplotypes, in order to generate the LD structure observed for this sample. We randomly specified the crossover locations such that the size of the steps in the LD maps was proportional to the likelihood of a recombination event. In other words, larger steps resulted in a higher chance of a crossover. For each crossover, we randomly picked a haplotype from the d haplotype pool and recorded the segment corresponding to the region on the disease harboring haplotype that was recombined.

The founder pools containing 5 independently selected mutations were simply a collection of 5 founder pools, each pool consisting of a distinct mutation.

3.2.11 Power Calculations to Evaluate the Performance of the Haplotype Sharing Measures, Summary Statistics, Clustering Algorithm, and Traditional Approaches

To calculate power of the haplotype sharing measures and accompanying summary statistics, clustering algorithm, and traditional approaches, for each of the mutational

models (i.e. 1 disease mutation inherited across 20 generations, 1 mutation / 100 generations, 5 mutations / 20 generations, and 5 mutations / 100 generations) we simulated 100 data sets using the BRCA1 and PHB gene region SNP sets discussed in Section 3.2.10. Then for each of the simulated data sets, we computed all single marker and haplotype association tests as described in Sections 3.2.4, 3.2.6, 3.2.7, and 3.2.8, across a selected subset of SNPs, designated as all of the SNPs in BRCA1 and PHB in addition to 5 SNPs up and downstream of these genes. For BRCA1, this subset started at rs8076790 (bp location 38,408,126) and ended at rs11651341 (bp location 38,783,587) for a total size of 375.46 kilobases on 22 SNPs (Table 3.1). For PHB, the segment analyzed ranged from rs2584663 (bp location 44,823,146) to rs4794054 (bp location 44,887,097) which was smaller than the BRCA1 region at 63.95 kilobases across a smaller set of 16 SNPs (Table 3.2).

The SNPs within the BRCA1 and PHB sets served as the analysis focal points for the battery of association tests performed. At a given SNP, we carried out the allelic test (Section 3.2.6) and for the haplotype χ^2 test we positioned this SNP in the center of windows that were 3, 7, and 11 SNPs wide (i.e. 1, 3, and 5 SNPs to each side of the given SNP) and carried out the "Delete" and "Collapse" methods. We used these same windows for the haplotype sharing measures based on fixed windows (Section 3.2.2), and these scores were $\log_{10}(\text{CHSS}_{w_1}^{w_2})$, $\text{Match}_{w_1}^{w_2}$, $\text{Length}_{w_1}^{w_2}$, and $\text{Count}_{w_1}^{w_2}$. This given SNP was also the reference marker for the scores generated from the reference marker approach (Section 3.2.1), which were $\log_{10}(\text{CHSS}^r)$, Length^r, and Count^r. To evaluate the Length reference marker and fixed window measures, we defined the physical positions of all of the SNPs in the data sets per the NCBI B36 assembly which corresponded to dbSNP b126. Subsequently, the clustering algorithm (Section 3.2.8) employed all of these reference marker and fixed window scores at this SNP. The threshold scores (Section 3.2.3) were computed for all of the reference marker and fixed window scores and were

based on 6 percentiles: 75%, 90%, 95%, 99%, 99.5%, and 99.9%. Lastly, for each score we calculated 3 summary statistics (Section 3.2.4), Sum^{Case}, Sum^{Diff}, and M, and estimated their statistical significance by setting the number of permutations, B, at 5,000 and then evaluating the permutation p-value in Equation 3.17. To compute M, we specified μ as 0.001 since we simulated the data sets under a disease prevalence K of 0.001.

In the $R \times 2$ contingency tables of the clustering algorithm, we enforced cluster sizes (i.e. row totals) to be at least 10 and we also compared the performance of the clustering algorithm when there was no cluster size restriction. This row total criterion was utilized in all 3 of the approaches of the $R \times 2$ clustering algorithm ("No Regrouping", "Regrouping", and "Small Cluster Row"). The same series of percentiles (75%, 90%, 95%, 99%, 99.5%, and 99.9%) utilized for the threshold scores was also employed in the $R \times 2$ and 2×2 clustering algorithm. Furthermore, for the 2×2 clustering algorithm, both 1- and 2-sided tests were performed.

At each of the 22 and 16 SNPs in the BRCA1 and PHB sets, respectively, we computed all of the above-mentioned test statistics and p-values across all 100 simulated data sets. For a given test and SNP, we calculated the power to be the proportion of times out of 100 that the test was significant at the 0.05 alpha level.

We adjusted the power estimates for multiple testing across the 22 and 16 SNPs by way of Bonferroni and an empirical method. Specifically, for the Bonferroni adjustment we determined the minimum p-value for a given test throughout all of the referent SNPs and for a given simulated data set. We then assessed the Bonferroni adjusted power by computing the proportion of times out of 100 the minimum p-value was significant at the 0.0023 (= 0.05 / 22) or 0.0031 (= 0.05 / 16) alpha level for the BRCA1 or PHB SNP sets, respectively. Additionally, we empirically controlled the overall type I error by first simulating 2,500 data sets under the null model that no association existed

between case-control status and the genotypes simulated at the putative disease locus. Then at a particular test and simulated null data set, we found the minimum p-value amongst the referent SNPs, resulting in an empirical null distribution of minimum p-values based on 2,500 minimum p-values. From this null distribution, we located the minimum p-value that demarcated the smallest 5% of the minimum p-values and designated this as the empirically determined significance cutoff that controlled the overall type I error rate at 5%. The location of this empirical threshold corresponded to the 125-th smallest minimum p-values. We note that these multiple testing methods corrected for the multiple tests conducted amongst the SNPs and did not adjust for the array of various types of association tests we performed.

We investigated the power of removing rare SNPs as well as SNPs in strong LD with each other. Specifically, one filter removed SNPs that had a MAF of 2% or lower and another filter excluded SNPs with pairwise r^2 of 1% or more. We calculated r^2 in PLINK (Purcell, 2007; Purcell et al., 2007) as the squared correlation based on genotypic allele counts. For each of the 100 BRCA1 and PHB simulated data sets, we analyzed the set of SNPs in 4 different ways: 1) without any filters, which included all SNPs; 2) 2% MAF filter applied; 3) 1% r^2 filter applied; 4) both the 2% MAF and 1% r^2 filters applied. We used PLINK's data management capabilities to exclude SNPs that did not meet the 2% MAF and 1% r^2 criteria.

We computed power in the smaller pruned BRCA1 and PHB data sets in the same manner as that described for the full data sets. The referent SNPs in the filtered data sets were the same 22 and 16 referent SNPs chosen previously from the full BRCA1 and PHB data sets, respectively, less any SNPs that were excluded due to the 2% MAF and/or 1% r^2 filter, depending on the exclusion scenario. Consequently, the Bonferroni and empirical corrections were based on the smaller number of analyzed SNPs.

Hardy-Weinberg equilibrium (HWE) was assumed throughout the power analyses.

3.2.12 Computational Aspects and Complexity

To offer the reader an appreciation of the magnitude of the tests performed for the power calculations across the group of SNPs about BRCA1 and PHB and the computational requirements this entailed both in collecting and summarizing the data, we begin by describing the tests conducted at a particular SNP and a given simulated data set:

1. Allelic test
2. Haplotype χ^2, Delete and Collapse methods, 3 windows each $\Rightarrow$ **6 tests**
3. Summing / Permutation tests
   (a) 3 reference marker scores
   (b) 4 fixed window scores, 3 windows each $\Rightarrow$ 12 scores
   (c) For each of the reference marker and fixed window scores, there was a non-threshold score plus binary and ratio threshold scores with 6 percentiles each $\Rightarrow$ 13 scores
   (d) For each score, there were 3 summary measures $\Rightarrow$ 15 (= 3 + 12) primary scores $\times$ 13 sub-scores $\times$ 3 summary statistics = **585 total tests**
4. $R \times 2$ clustering algorithm
   (a) 3 reference marker scores
   (b) 3 fixed window scores, 3 windows each $\Rightarrow$ 9 scores
   (c) For each of the scores, we defined 6 percentiles to compute thresholds so as to cluster the haplotypes
   (d) For each percentile, there were 12 test types, i.e. no minimum cluster size / cluster size of 10, No Regrouping / Regrouping / Small Cluster Row, and

Keep / Delete ⇒ 12 (= 3 + 9) scores × 6 percentiles × 12 test types = **864 total tests**

5. 2 × 2 clustering algorithm

   (a) 3 reference marker scores

   (b) 3 fixed window scores, 3 windows each ⇒ 9 scores

   (c) For each of the scores, we defined 6 percentiles to compute thresholds so as to cluster the haplotypes

   (d) For each percentile, we conducted both 1- and 2-sided tests ⇒ 12 (= 3 + 9) scores × 6 percentiles × 2 = **144 total tests**

6. **Grand total number of distinct tests per SNP = 1,600**

Moreover, for the full data sets of BRCA1 and PHB we analyzed 22 and 16 SNPs, respectively, so there were 35,200 (= 1,600 tests × 22 SNPs) and 25,600 (= 1,600 tests × 16 SNPs) tests performed. We further applied the 3 MAF and LD exclusion criteria as described in Section 3.2.11, resulting in the following number of tests conducted: 1) 2% MAF filter — there were 21 (BRCA1 set) and 15 (PHB set) SNPs for analysis, contributing 33,600 and 24,000 tests; 2) 1% r^2 filter — there were 2 referent SNPs for each gene set, adding 3,200 tests each; 3) 2% MAF and 1% r^2 filters — there was 1 referent SNP for each gene set, furnishing 1,600 tests each. Thus, for BRCA1 and PHB we conducted 73,600 and 54,400 tests, respectively, for each of the 4 mutational models (i.e. 1 disease mutation inherited across 20 generations, 1 mutation / 100 generations, 5 mutations / 20 generations, and 5 mutations / 100 generations), bringing the number of tests to 294,400 and 217,600. Lastly, to compute the power we iterated through these tests 100 times, for a grand total of 294,400,000 and 217,600,000 tests carried out which included BRCA1 and PHB, all mutational models, and all SNP exclusion criteria.

Data preprocessing, the haplotype pairing scores (Sections 3.2.1 and 3.2.2), threshold scores (Section 3.2.3), summary statistics (Section 3.2.4), permutation tests (Section 3.2.5), allelic test (Section 3.2.6), haplotype χ^2 test (Section 3.2.7), and clustering algorithm (Section 3.2.8) were coded in C. We used a select set of subroutines from the Numerical Recipes in C UNIX/Linux Version 2.10 (Press et al., 2002; software and license obtained from www.nr.com), including "ran2" to generate uniform random deviates, "select" to return the k-th smallest value from a given array of values, "sort2" to sort an array into ascending order using Quicksort while making the corresponding rearrangements of another array, and "gammq" to compute p-values from the χ^2 distribution. Appropriate changes were made to the source code of these subroutines so as to conform to the architecture of our code. There were a total of 19,644 lines of code spanning 384 single-sided printed pages (an uncompressed text file of almost 1 megabyte in size), which consisted of current and older versions of original subroutines, the Numerical Recipes code, and detailed comments throughout.

We coded the simulation engines in both C and R, which generated the simulated data sets as discussed in Section 3.2.10. The C code covered the probability sampling of haplotypes and the R code constituted the construction of the founder pools.

In order to compute the power of the multitude of tests described above for the BRCA1 and PHB gene sets, across all combinations of the 4 MAF/LD exclusion criteria and 4 mutational models (Section 3.2.11), we had the computational burden of processing 400 referent SNPs. For each of the simulated data sets, the 5,000 permutations to estimate the statistical significance of the summary statistics (Section 3.2.4) was incredibly time consuming, initially carried out serially on UNC's Emerald, a 850-processor Beowulf Linux cluster. In lieu of running these jobs for several months upon end, in the C code we parallelized the permutation step by coding this with Message Passing Interface (MPI) so as to take advantage of the parallel computing environment

offered by UNC's Topsail, a 4,160-processor Dell Linux cluster that was ranked by TOP500 (www.top500.org) as the 87-th fastest publicly known supercomputer in the world.

As an example, to analyze a particular SNP within the PHB gene set, on Topsail it took 24 processors running in parallel almost 5 hours to complete 100 simulated data sets. This was equivalent to over 4 and a half days of CPU time. On the other hand, analyzing 2,500 null data sets to determine the null distribution of minimum p-values for the empirically evaluated multiple test adjustments was more demanding. For example, at a single SNP, 32 parallel processors ran for over 3 days, equivalent to over 100 days of CPU time.

We wrote numerous Bash scripts that aided in automatically deploying the massive number of jobs on Topsail as well as organizing the results, approximately 17 gigabytes worth. In summarizing the results, we wrote R code that arranged the power estimates in tables formatted in LaTeX. Across BRCA1 and PHB, all mutational models, and all SNP exclusion criteria, there were 1,632 one-sided landscape pages of results.

Finally, the LD displays in the Results (Section 3.3) were produced using Haploview version 4.1 (Barrett et al., 2005).

3.3 Results

Upon subsetting the genome-wide SNP data from the consensus set of HumanHap550 version 1 and 3 platforms, there were 14,109 SNPs originating from chromosome 17. We further restricted the subjects to Caucasian; there were a total of 3,172 Caucasians (58.27% of the sample) of which there were 1,579 and 1,593 from versions 1 and 3, respectively. We removed 87 (2.74% of the Caucasians) individuals who were likely related based on their computed $\hat{\pi}$, leaving 3,085 unrelated and independent Caucasians.

None of these subjects were missing more than 5% of their genotypes and so none were removed for low genotyping. However, 214 SNPs (1.52%) were missing genotypes at a rate of more than 5%, thus we eliminated them from further analysis, resulting in a total of 13,895 chromosome 17 SNPs. Lastly, although we computed the MAFs, we allowed rare SNPs to remain in the data set.

Phasing of the genotypes in fastPHASE required almost 26 days (618 hours) using one processor and a maximum of 296 and 313 megabytes of random access memory and swap space, respectively, on UNC's Emerald, a 850-processor Beowulf Linux cluster. The 6,170 inferred and independent haplotypes (each of the 3,085 subjects contributed 2 haplotypes) served as the "d haplotype pool" from which we sampled to simulate the data sets for the power analyses, as described in Section 3.2.10.

The SNP names, physical positions, description of locations, alleles, and computed MAFs of the set of SNPs that we designated as the referent SNPs for the power analyses are presented in Tables 3.1 and 3.2 for BRCA1 and PHB, respectively. In the BRCA1 set, there were 5 SNPs that were relatively rare as their MAFs were less than 10% (rs775990, rs8176225, rs3737559, rs4793211, rs8078799) and rs8176225 was incredibly uncommon with a MAF of 0.1%, located in the intron of BRCA1. On the other hand, more than half of the SNPs' MAFs were about 30%. In the PHB set, there were 3 SNPs that had MAFs of less than 10% (rs8065814, rs8066722, rs2277636) in which rs2277636 was quite infrequent (MAF = 0.9%), positioned in the intron of PHB. Conversely, the majority of SNPs had MAFs greater than 30%, of which a handful were above 40%.

Figures 3.1 and 3.2 provide a graphical display of the LD patterns observed in the BRCA1 and PHB data sets, respectively, that covered an approximate 3 megabase segment roughly centered about the genes of interest. In the BRCA1 data set (Figure 3.1), there were distinctive LD blocks of varying sizes throughout the region, and BRCA1 exhibited the largest conserved area. On the contrary, in the PHB data set (Figure

TABLE 3.1: Characterization of the 22 SNPs chosen within (12 SNPs) and surrounding the BRCA1 gene (5 SNPs up and downstream). These SNPs served as the referent locations for the subsequent power analyses.

Index	dbSNP ID[a]	Nucleotide Position[b]	Location	Alleles[c]	MAF[d]
1	rs8076790	38,408,126	RPL27, intron	C / T	0.200
2	rs775990	38,412,059	IFI35	C / T	0.059
3	rs382571	38,425,007	VAT1, intron	G / A	0.183
4	rs9911630	38,441,868		G / A	0.355
5	rs11657053	38,444,655		T / G	0.336
6	rs8176273	38,465,179	BRCA1, intron	C / T	0.333
7	rs8176265	38,467,522	BRCA1, intron	A / G	0.333
8	rs8176257	38,469,731	BRCA1, intron	A / C	0.272
9	rs8176225	38,475,122	BRCA1, intron	T / G	0.001
10	rs1799966	38,476,620	BRCA1, intron	G / A	0.333
11	rs3737559	38,487,830	BRCA1, intron	A / G	0.077
12	rs1060915	38,487,996	BRCA1, intron	C / T	0.335
13	rs16942	38,497,526	BRCA1, intron	G / A	0.335
14	rs799917	38,498,462	BRCA1, intron	T / C	0.357
15	rs16940	38,498,763	BRCA1, intron	C / T	0.333
16	rs1799949	38,498,992	BRCA1, intron	T / C	0.269
17	rs799923	38,505,457	BRCA1, intron	A / G	0.229
18	rs4793211	38,552,381		C / T	0.017
19	rs9646417	38,779,779		A / G	0.337
20	rs8078799	38,782,474		A / G	0.010
21	rs4793230	38,782,929		C / A	0.352
22	rs11651341	38,783,587		C / T	0.339

[a]dbSNP: www.ncbi.nlm.nih.gov/projects/SNP
[b]Nucleotide positions based on NCBI B36 assembly, dbSNP b126
[c]Minor / major allele
[d]Minor allele frequency
[e]Number of non-missing alleles
RPL27: ribosomal protein L27
IFI35: interferon-induced protein 35
VAT1: vesicle amine transport protein 1 homolog (T. californica)
BRCA1: breast cancer susceptibility gene 1, early onset

TABLE 3.2: Characterization of the 16 SNPs chosen within (6 SNPs) and surrounding the PHB gene (5 SNPs up and downstream). These SNPs served as the referent locations for the subsequent power analyses.

Index	dbSNP ID[a]	Nucleotide Position[b]	Location	Alleles[c]	MAF[d]
1	rs2584663	44,823,146		T / C	0.356
2	rs8065814	44,825,719		C / T	0.077
3	rs8066722	44,830,006		A / G	0.078
4	rs2197159	44,832,634		C / T	0.328
5	rs4987082	44,836,373		G / A	0.410
6	rs1049620	44,836,513	PHB, UTR	A / G	0.197
7	rs2898883	44,837,952	PHB, intron	A / G	0.305
8	rs2233669	44,839,002	PHB, intron	G / A	0.435
9	rs935129	44,841,015	PHB, intron	A / G	0.305
10	rs7502499	44,845,101	PHB, intron	A / G	0.302
11	rs2277636	44,847,176	PHB, intron	T / C	0.009
12	rs7222591	44,862,018		T / G	0.406
13	rs2119930	44,869,038		C / A	0.416
14	rs2584684	44,875,052		G / A	0.125
15	rs2584681	44,884,745		T / C	0.269
16	rs4794054	44,887,097		T / G	0.114

[a]dbSNP: www.ncbi.nlm.nih.gov/projects/SNP
[b]Nucleotide positions based on NCBI B36 assembly, dbSNP b126
[c]Minor / major allele
[d]Minor allele frequency
[e]Number of non-missing alleles
PHB: prohibitin
UTR: untranslated region

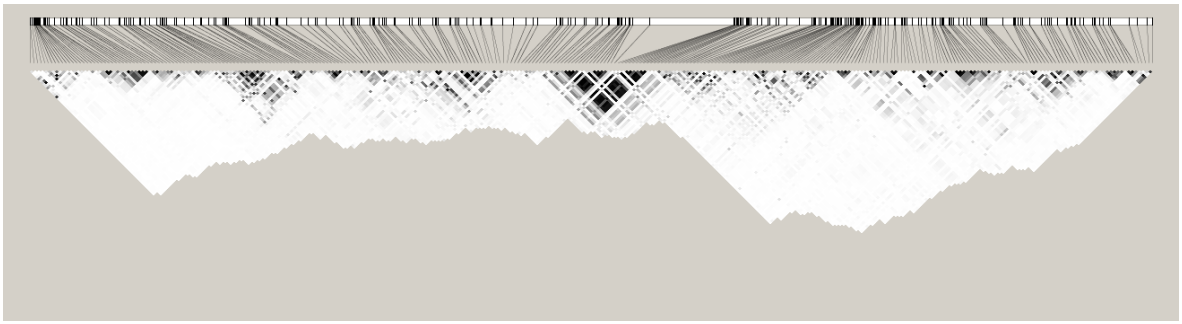

FIGURE 3.1: LD plot (r^2) of the 3.03 megabase region consisting of the BRCA1 gene approximately centered within flanking segments. There were a total of 314 SNPs, of which 12 resided in BRCA1. White represents $r^2 = 0$, shades of grey $0 < r^2 < 1$, and black $r^2 = 1$.

FIGURE 3.2: LD plot (r^2) of the 3.03 megabase region consisting of the PHB gene approximately centered within flanking segments. There were a total of 486 SNPs, of which 6 resided in PHB. White represents $r^2 = 0$, shades of grey $0 < r^2 < 1$, and black $r^2 = 1$.

3.2), the LD blocks were less apparent though indeed present, suggesting a greater degree of diversity along this portion of the chromosome. PHB, located to the right of a medium-sized LD block and approximately in the center of Figure 3.2, was barely noticeable in addition to much weaker r^2 values computed within the gene, compared to BRCA1.

Figures 3.3 and 3.4 show a closer view of BRCA1 and PHB, respectively, that emphasizes the difference between these genes as well as the areas directly up and downstream of them. There were 12 SNPs residing in BRCA1 (SNPs 6 through 17 in

Figure 3.3) that constituted a large number of high r^2 values (black squares). The rarest SNP, rs8176225, was not correlated to any of the SNPs in BRCA1 ($r^2 = 0$ throughout, white squares). The other uncommon SNP, rs3737559, though more prevalent than rs8176225 exhibited weak correlations with all of the SNPs in BRCA1. Downstream of BRCA1 (SNPs 18 through 22), r^2 values remained strong with the exception of 2 less common SNPs, rs4793211 and rs8078799. Upstream of BRCA1 (SNPs 1 through 5), the correlations appeared to taper off.

In contrast to BRCA1, PHB contained half as many SNPs (SNPs 6 through 11 in Figure 3.4), likely owing to its smaller physical size (BRCA1: 81.16 kilobases; PHB: 10.82 kilobases). There was only one high pairwise r^2 value, whereas the other correlations were largely moderate (shades of grey). The correlations became weaker downstream of PHB (SNPs 12 through 16), while on the contrary the correlations were moderate at best upstream of PHB (SNPs 1 through 5).

These contrasting gene regions of BRCA1 and PHB allowed us to compare the performance of the haplotype analysis techniques between hypothetical disease harboring chromosomal segments that differed in LD and physical size.

Table 3.3 contains the results of the power analysis conducted on the BRCA1 data sets, in which single founder mutation events were simulated and followed throughout 100 generations, for the $\log_{10}$(CHSS), Length, and Count reference marker scores and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M at each of the 22 selected SNPs within and surrounding the BRCA1 gene. The Bonferroni and empirically adjusted powers of the $\log_{10}$(CHSS) reference marker score as well as the powers computed at each of the 22 SNPs were much greater than the Length and Count scores, for which their powers were about the same after controlling for multiple tests. Sum^{Cas} and M had nearly the same power across all 22 referent SNPs and after adjustment, while Sum^{Diff} was markedly consistently lower. Interestingly, for Sum^{Cas} and M based

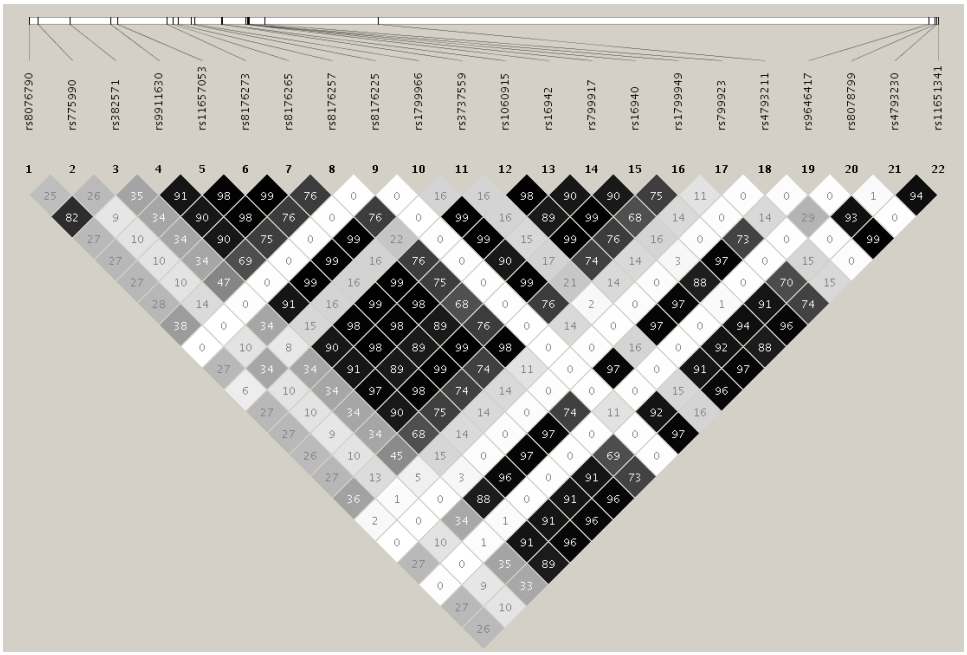

FIGURE 3.3: LD plot (r^2) of the BRCA1 gene surrounded by 5 SNPs up and downstream, for a total size of 375.5 kilobases and 22 referent SNPs. White represents $r^2 = 0$, shades of grey $0 < r^2 < 1$, and black $r^2 = 1$.

on $\log_{10}$(CHSS), the powers at the SNPs within BRCA1 (SNPs 6 through 17) were overall higher than at the 5 locations upstream (SNPs 1 through 5), in contrast to the 5 SNPs downstream (SNPs 18 through 22) for which the powers steadily climbed to a maximum of 0.96.

Instead of one founder mutation for a given simulated data set, five founder mutations were created and the disease allele frequency was set slightly higher at $f_D = 0.02$ in lieu of $f_D = 0.0125$ for the one mutation models. All of the other model parameters remained the same (i.e. disease prevalence $K = 0.001$, recessive genetic mode-of-inheritance risk model, and genotype relative risk $GRR = 750$). Table 3.4 contains the five mutation model results. All of the same patterns and features discussed for Table 3.3 also held for Table 3.4. Specifying the same model parameters for the five mutation model as for the one mutation model returned overall lower powers (data not

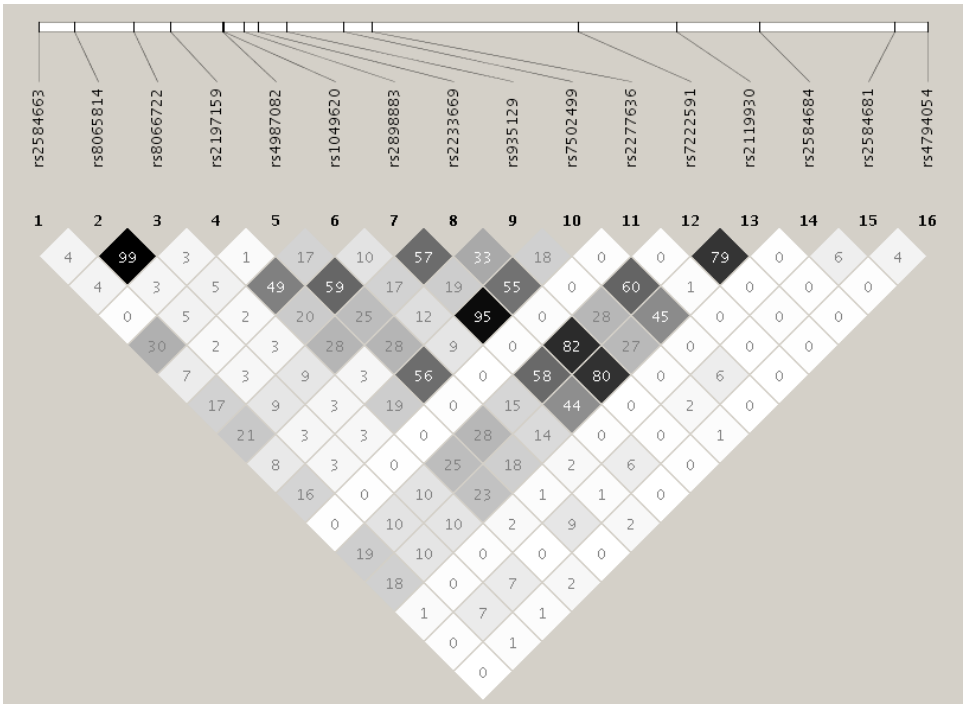

FIGURE 3.4: LD plot (r^2) of the PHB gene surrounded by 5 SNPs up and downstream, for a total size of 64.0 kilobases and 16 referent SNPs. White represents $r^2 = 0$, shades of grey $0 < r^2 < 1$, and black $r^2 = 1$.

shown), suggesting that the $\log_{10}$(CHSS), Length, and Count reference marker scores and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M were not as powerful in detecting disease associations when multiple independent mutations were present in a given founder pool.

For the remaining analyses (Tables 3.5 through 3.12) based on causal mutations in BRCA1, we simulated a single founder mutation event for each simulated data set.

TABLE 3.3: Power of the $\log_{10}$(CHSS), Length, and Count reference marker scores and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M at each of the 22 selected SNPs within and surrounding the BRCA1 gene. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

	Summary	Marker Positions																						
Method	Statistic	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Bonf.	Emp.
$\log_{10}$(CHSS)	Sum^{Cas}	0.75	0.71	0.74	0.87	0.88	0.88	0.87	0.84	0.89	0.88	0.83	0.89	0.90	0.90	0.90	0.87	0.90	0.93	0.94	0.95	0.95	0.96	0.74	0.91
	Sum^{Diff}	0.51	0.49	0.51	0.71	0.70	0.70	0.71	0.67	0.69	0.70	0.62	0.76	0.78	0.76	0.76	0.72	0.79	0.75	0.81	0.79	0.81	0.81	0.58	0.81
	M	0.75	0.71	0.74	0.87	0.88	0.88	0.87	0.84	0.89	0.88	0.83	0.90	0.90	0.90	0.90	0.87	0.90	0.93	0.94	0.95	0.95	0.96	0.74	0.91
Length	Sum^{Cas}	0.44	0.40	0.43	0.43	0.43	0.42	0.42	0.41	0.41	0.41	0.39	0.43	0.43	0.43	0.43	0.43	0.69	0.52	0.52	0.51	0.51	0.50	0.30	0.65
	Sum^{Diff}	0.33	0.34	0.32	0.33	0.33	0.32	0.33	0.32	0.32	0.32	0.32	0.35	0.34	0.35	0.35	0.35	0.54	0.37	0.37	0.38	0.37	0.37	0.22	0.51
	M	0.44	0.40	0.42	0.43	0.43	0.42	0.42	0.41	0.41	0.41	0.39	0.43	0.43	0.43	0.43	0.43	0.69	0.52	0.52	0.51	0.51	0.50	0.30	0.65
Count	Sum^{Cas}	0.49	0.48	0.49	0.49	0.50	0.50	0.50	0.49	0.50	0.50	0.49	0.51	0.52	0.51	0.50	0.50	0.72	0.65	0.63	0.65	0.65	0.64	0.33	0.65
	Sum^{Diff}	0.34	0.35	0.35	0.35	0.35	0.34	0.35	0.35	0.35	0.35	0.34	0.36	0.37	0.36	0.36	0.36	0.58	0.44	0.44	0.44	0.44	0.45	0.23	0.45
	M	0.49	0.48	0.49	0.49	0.50	0.50	0.50	0.49	0.50	0.50	0.49	0.51	0.52	0.51	0.50	0.50	0.72	0.65	0.63	0.65	0.65	0.64	0.33	0.65

TABLE 3.4: Power of the $\log_{10}$(CHSS), Length, and Count reference marker scores and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M at each of the 22 selected SNPs within and surrounding the BRCA1 gene. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. Five independent founder mutations carried throughout 100 generations were simulated.

	Summary	Marker Positions																						
Method	Statistic	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Bonf.	Emp.
$\log_{10}$(CHSS)	Sum^{Cas}	0.72	0.68	0.71	0.83	0.84	0.84	0.84	0.79	0.84	0.83	0.75	0.84	0.85	0.82	0.83	0.80	0.91	0.91	0.95	0.94	0.94	0.95	0.79	0.95
	Sum^{Diff}	0.48	0.44	0.46	0.55	0.58	0.56	0.56	0.53	0.58	0.56	0.51	0.61	0.59	0.60	0.59	0.55	0.75	0.64	0.65	0.65	0.67	0.66	0.53	0.74
	M	0.72	0.68	0.72	0.82	0.84	0.84	0.84	0.79	0.85	0.83	0.75	0.84	0.85	0.82	0.84	0.80	0.91	0.91	0.95	0.94	0.94	0.95	0.79	0.95
Length	Sum^{Cas}	0.35	0.37	0.37	0.41	0.37	0.38	0.40	0.36	0.36	0.37	0.37	0.40	0.39	0.40	0.40	0.42	0.59	0.40	0.36	0.39	0.39	0.39	0.27	0.58
	Sum^{Diff}	0.25	0.28	0.27	0.26	0.26	0.25	0.25	0.25	0.25	0.25	0.25	0.26	0.28	0.25	0.27	0.27	0.41	0.27	0.29	0.28	0.27	0.28	0.15	0.41
	M	0.35	0.37	0.37	0.41	0.37	0.38	0.40	0.36	0.36	0.37	0.37	0.40	0.39	0.40	0.40	0.42	0.59	0.40	0.36	0.39	0.39	0.39	0.27	0.58
Count	Sum^{Cas}	0.40	0.41	0.41	0.42	0.41	0.41	0.41	0.40	0.41	0.41	0.40	0.42	0.41	0.43	0.41	0.42	0.60	0.50	0.47	0.52	0.49	0.49	0.30	0.55
	Sum^{Diff}	0.31	0.32	0.33	0.29	0.30	0.30	0.29	0.27	0.30	0.28	0.30	0.29	0.30	0.29	0.29	0.29	0.47	0.35	0.35	0.32	0.33	0.35	0.15	0.39
	M	0.40	0.41	0.41	0.42	0.41	0.41	0.41	0.40	0.41	0.41	0.40	0.42	0.41	0.43	0.41	0.42	0.60	0.50	0.47	0.52	0.49	0.49	0.30	0.55

The power of the binary and ratio threshold scores, computed from the $\log_{10}$(CHSS) reference marker measures, and Sum^{Cas}, Sum^{Diff}, and M are shown in Table 3.5. For both the binary and ratio threshold scores, as the percentiles increased from 75% to 99.9%, the powers increased as well. The binary scores outperformed the ratio scores; the binary scores achieved a power of 1.00 at a 99% threshold, empirically adjusted, whereas the maximum empirically adjusted power from the ratio scores was 0.92 for a stringent 99.9% threshold. As the threshold percentiles increased, the differences in power between the $\text{Sum}^{\text{Cas}}/M$ and Sum^{Diff} summary statistics became greater (Sum^{Diff} having lower powers) and then gradually became less for the higher percentiles (99%, 99.5%, and 99.9%), with the Sum^{Cas} and M having equivalent powers. For the binary scores, we did not observe large differences in power depending on the location of the reference marker (e.g. within BRCA1 versus outside), with the exception of the first 5 SNPs (SNPs 1 through 5) when a 95% threshold was used. On the other hand, for the ratio scores, beginning at about the 99% threshold, the earlier SNPs (SNPs 1 through 5) tended to have lower power compared to the powers for SNPs within BRCA1 and powers at the posterior SNPs (SNPs 18 through 22) were higher than the powers for SNPs within BRCA1, as we saw before in Tables 3.3 and 3.4.

TABLE 3.5: Power of the $\log_{10}$(CHSS) reference marker threshold scores (binary and ratio) and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M at each of the 22 selected SNPs within and surrounding the BRCA1 gene. The 75%, 90%, 95%, 99%, 99.5%, and 99.9% thresholds were considered. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

Threshold (%)	Summary Statistic	Marker Positions 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Bonf.	Emp.
Binary																									
75	Sum^{Cas}	0.24	0.25	0.25	0.30	0.31	0.28	0.29	0.27	0.29	0.28	0.28	0.29	0.30	0.29	0.31	0.30	0.29	0.31	0.32	0.31	0.30	0.32	0.17	0.34
	Sum^{Diff}	0.23	0.24	0.23	0.25	0.25	0.25	0.24	0.24	0.24	0.24	0.22	0.24	0.26	0.25	0.24	0.25	0.24	0.26	0.26	0.26	0.26	0.26	0.14	0.31
	M	0.24	0.25	0.25	0.30	0.31	0.28	0.29	0.27	0.29	0.28	0.28	0.29	0.30	0.29	0.31	0.30	0.29	0.31	0.32	0.31	0.30	0.32	0.17	0.34
90	Sum^{Cas}	0.39	0.38	0.43	0.40	0.39	0.39	0.39	0.39	0.40	0.39	0.39	0.40	0.39	0.39	0.39	0.39	0.45	0.46	0.46	0.46	0.46	0.46	0.26	0.49
	Sum^{Diff}	0.27	0.22	0.24	0.32	0.33	0.34	0.33	0.34	0.35	0.34	0.33	0.33	0.33	0.34	0.33	0.33	0.30	0.33	0.33	0.35	0.33	0.34	0.15	0.33
	M	0.39	0.38	0.43	0.40	0.39	0.39	0.39	0.39	0.40	0.39	0.39	0.40	0.39	0.39	0.39	0.39	0.45	0.46	0.46	0.46	0.46	0.46	0.26	0.49
95	Sum^{Cas}	0.70	0.74	0.71	0.62	0.62	0.63	0.62	0.62	0.62	0.62	0.64	0.62	0.61	0.61	0.61	0.62	0.61	0.61	0.61	0.62	0.62	0.62	0.45	0.62
	Sum^{Diff}	0.49	0.52	0.50	0.48	0.48	0.47	0.47	0.48	0.50	0.51	0.48	0.45	0.46	0.47	0.46	0.46	0.48	0.48	0.47	0.47	0.48	0.47	0.28	0.45
	M	0.70	0.74	0.71	0.62	0.62	0.63	0.62	0.62	0.63	0.62	0.64	0.62	0.61	0.61	0.61	0.62	0.61	0.61	0.61	0.62	0.62	0.62	0.45	0.62
99	Sum^{Cas}	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00
	Sum^{Diff}	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.88	0.96
	M	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00
99.5	Sum^{Cas}	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00
	Sum^{Diff}	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.98
	M	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00
99.9	Sum^{Cas}	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.99
	Sum^{Diff}	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.99
	M	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.99
Ratio																									
75	Sum^{Cas}	0.21	0.21	0.20	0.24	0.24	0.24	0.24	0.21	0.22	0.23	0.23	0.23	0.24	0.24	0.25	0.24	0.25	0.26	0.29	0.28	0.29	0.30	0.07	0.31
	Sum^{Diff}	0.21	0.21	0.19	0.22	0.22	0.22	0.23	0.21	0.21	0.21	0.22	0.21	0.22	0.20	0.23	0.20	0.26	0.22	0.25	0.23	0.24	0.26	0.07	0.30
	M	0.21	0.21	0.20	0.24	0.24	0.24	0.24	0.21	0.22	0.23	0.23	0.23	0.24	0.24	0.25	0.24	0.25	0.26	0.29	0.28	0.29	0.30	0.07	0.31
90	Sum^{Cas}	0.25	0.25	0.25	0.27	0.28	0.27	0.29	0.27	0.28	0.28	0.26	0.29	0.29	0.25	0.27	0.26	0.34	0.29	0.33	0.33	0.33	0.33	0.09	0.36
	Sum^{Diff}	0.25	0.25	0.24	0.23	0.26	0.25	0.26	0.24	0.23	0.23	0.23	0.24	0.24	0.21	0.20	0.20	0.27	0.24	0.28	0.28	0.27	0.28	0.08	0.29
	M	0.25	0.25	0.25	0.27	0.28	0.27	0.29	0.27	0.28	0.28	0.26	0.29	0.29	0.25	0.27	0.26	0.34	0.29	0.33	0.33	0.33	0.33	0.09	0.36
95	Sum^{Cas}	0.26	0.27	0.28	0.31	0.30	0.30	0.30	0.29	0.30	0.30	0.29	0.33	0.33	0.31	0.31	0.31	0.41	0.36	0.40	0.39	0.40	0.42	0.19	0.47
	Sum^{Diff}	0.25	0.23	0.25	0.25	0.27	0.26	0.26	0.25	0.24	0.25	0.27	0.25	0.24	0.21	0.21	0.20	0.31	0.26	0.30	0.29	0.29	0.30	0.11	0.39
	M	0.26	0.27	0.28	0.31	0.30	0.30	0.30	0.29	0.30	0.30	0.29	0.33	0.33	0.31	0.31	0.31	0.41	0.36	0.40	0.39	0.40	0.42	0.19	0.47
99	Sum^{Cas}	0.37	0.39	0.38	0.49	0.51	0.52	0.51	0.45	0.48	0.50	0.41	0.56	0.55	0.54	0.54	0.50	0.69	0.62	0.66	0.65	0.65	0.67	0.37	0.69
	Sum^{Diff}	0.31	0.32	0.33	0.30	0.30	0.30	0.31	0.32	0.31	0.31	0.30	0.36	0.36	0.36	0.36	0.34	0.49	0.43	0.50	0.48	0.50	0.49	0.22	0.58
	M	0.37	0.39	0.38	0.49	0.51	0.52	0.51	0.45	0.48	0.50	0.41	0.56	0.55	0.55	0.55	0.50	0.69	0.62	0.66	0.65	0.65	0.67	0.37	0.69
99.5	Sum^{Cas}	0.48	0.50	0.47	0.64	0.63	0.64	0.63	0.58	0.61	0.61	0.55	0.67	0.69	0.70	0.68	0.64	0.79	0.77	0.77	0.76	0.77	0.78	0.54	0.81
	Sum^{Diff}	0.36	0.38	0.37	0.43	0.43	0.45	0.45	0.41	0.42	0.43	0.38	0.48	0.49	0.46	0.46	0.43	0.59	0.54	0.54	0.55	0.57	0.54	0.34	0.68
	M	0.48	0.50	0.47	0.64	0.63	0.64	0.63	0.58	0.61	0.61	0.55	0.67	0.69	0.70	0.69	0.64	0.79	0.77	0.77	0.76	0.77	0.78	0.54	0.81
99.9	Sum^{Cas}	0.72	0.67	0.72	0.85	0.85	0.83	0.83	0.83	0.83	0.83	0.80	0.87	0.87	0.87	0.87	0.86	0.88	0.90	0.90	0.90	0.90	0.91	0.70	0.92
	Sum^{Diff}	0.48	0.48	0.49	0.66	0.68	0.67	0.68	0.63	0.66	0.66	0.61	0.72	0.74	0.73	0.72	0.70	0.75	0.74	0.76	0.75	0.77	0.78	0.56	0.83
	M	0.72	0.67	0.72	0.85	0.85	0.83	0.83	0.83	0.83	0.83	0.80	0.87	0.87	0.87	0.87	0.86	0.88	0.90	0.90	0.90	0.90	0.91	0.70	0.92

We determined the power of employing windows of fixed lengths in comparison to the reference marker approach, as discussed in Section 3.2.2. Table 3.6 presents the power of the $\log_{10}$(CHSS), Length, Count, and Match scores for fixed windows of sizes 3, 7, and 11 SNPs and accompanying summary statistics. None of the measures achieved the same or greater amount of power than their reference marker counterparts (i.e. the $\log_{10}$(CHSS), Length, and Count measures) that were demonstrated in Table 3.3. In fact, the highest empirically adjusted power attained by the window based scores was 0.50 by the $\log_{10}$(CHSS) and $\text{Sum}^{\text{Cas}}/M$, which was 0.41 lower than its reference marker analog at 0.91, also the highest power amongst all of the reference marker scores. The powers of the Length and Count fixed window scores were also well below their reference marker versions. The Match score almost performed as well as the $\log_{10}$(CHSS), its maximum occurring with the empirically adjusted power of Sum^{Diff} at 0.44. Across all of the measures, the smaller sized window of length 3 resulted in higher powers after controlling for multiple testing. Lastly, it did not appear that Sum^{Diff} was significantly less powerful than Sum^{Cas} and M, as we observed previously for the reference marker scores (Tables 3.3, 3.4, and 3.5).

TABLE 3.6: Power of the $\log_{10}$(CHSS), Length, Count, and Match scores for fixed windows of sizes 3, 7, and 11 SNPs and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M at each of the 22 selected SNPs within and surrounding the BRCA1 gene. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

Window Size	Summary Statistic	Marker Positions 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Bonf.	Emp.
$\log_{10}$(CHSS)																									
3	Sum^{Cas}	0.22	0.20	0.19	0.18	0.20	0.19	0.21	0.20	0.19	0.17	0.17	0.17	0.19	0.20	0.20	0.17	0.22	0.20	0.20	0.17	0.16	0.17	0.27	0.50
	Sum^{Diff}	0.20	0.19	0.16	0.17	0.20	0.19	0.20	0.18	0.17	0.16	0.16	0.16	0.19	0.21	0.19	0.15	0.13	0.13	0.19	0.17	0.16	0.17	0.22	0.46
	M	0.22	0.20	0.19	0.18	0.20	0.19	0.21	0.20	0.19	0.17	0.17	0.17	0.20	0.21	0.20	0.17	0.22	0.20	0.20	0.17	0.16	0.17	0.28	0.50
7	Sum^{Cas}	0.25	0.24	0.23	0.21	0.18	0.19	0.19	0.20	0.20	0.19	0.19	0.19	0.18	0.18	0.17	0.16	0.17	0.17	0.15	0.17	0.18	0.16	0.12	0.26
	Sum^{Diff}	0.24	0.20	0.21	0.19	0.16	0.17	0.19	0.18	0.18	0.18	0.18	0.17	0.16	0.18	0.16	0.15	0.16	0.16	0.15	0.15	0.18	0.16	0.12	0.26
	M	0.25	0.24	0.23	0.21	0.18	0.19	0.20	0.20	0.20	0.19	0.19	0.19	0.18	0.18	0.17	0.16	0.17	0.17	0.15	0.17	0.18	0.16	0.12	0.26
11	Sum^{Cas}	0.23	0.22	0.20	0.21	0.19	0.17	0.17	0.19	0.18	0.19	0.20	0.19	0.19	0.18	0.18	0.18	0.15	0.15	0.16	0.16	0.16	0.16	0.10	0.23
	Sum^{Diff}	0.22	0.21	0.18	0.17	0.18	0.15	0.14	0.18	0.19	0.20	0.18	0.18	0.18	0.16	0.16	0.16	0.16	0.15	0.17	0.15	0.15	0.16	0.08	0.22
	M	0.23	0.22	0.20	0.21	0.19	0.17	0.17	0.19	0.18	0.19	0.20	0.19	0.19	0.18	0.18	0.18	0.15	0.15	0.16	0.16	0.16	0.16	0.10	0.23
Length																									
3	Sum^{Cas}	0.19	0.15	0.12	0.14	0.21	0.22	0.16	0.19	0.19	0.17	0.17	0.20	0.21	0.14	0.14	0.21	0.15	0.18	0.17	0.17	0.19	0.18	0.14	0.37
	Sum^{Diff}	0.17	0.15	0.12	0.14	0.21	0.22	0.15	0.19	0.19	0.18	0.19	0.20	0.20	0.14	0.13	0.21	0.14	0.19	0.15	0.17	0.20	0.18	0.13	0.37
	M	0.19	0.15	0.12	0.14	0.21	0.22	0.16	0.19	0.19	0.17	0.17	0.20	0.21	0.14	0.14	0.21	0.15	0.18	0.17	0.17	0.19	0.18	0.14	0.37
7	Sum^{Cas}	0.14	0.15	0.15	0.17	0.17	0.17	0.22	0.19	0.18	0.18	0.18	0.18	0.18	0.22	0.17	0.19	0.18	0.18	0.20	0.19	0.18	0.17	0.06	0.31
	Sum^{Diff}	0.13	0.14	0.15	0.16	0.16	0.17	0.22	0.19	0.18	0.18	0.17	0.18	0.19	0.23	0.19	0.19	0.19	0.19	0.18	0.19	0.16	0.16	0.07	0.31
	M	0.14	0.15	0.15	0.17	0.17	0.17	0.22	0.19	0.18	0.18	0.18	0.18	0.18	0.22	0.17	0.19	0.18	0.18	0.20	0.19	0.18	0.17	0.06	0.31
11	Sum^{Cas}	0.15	0.14	0.14	0.17	0.17	0.17	0.17	0.18	0.19	0.17	0.18	0.22	0.20	0.23	0.21	0.20	0.21	0.18	0.21	0.20	0.21	0.23	0.06	0.21
	Sum^{Diff}	0.13	0.14	0.14	0.16	0.17	0.17	0.17	0.17	0.18	0.16	0.18	0.21	0.19	0.21	0.20	0.19	0.19	0.19	0.18	0.19	0.19	0.19	0.06	0.21
	M	0.15	0.14	0.14	0.17	0.17	0.17	0.17	0.18	0.19	0.17	0.18	0.22	0.20	0.23	0.21	0.20	0.21	0.18	0.21	0.20	0.21	0.23	0.06	0.21
Count																									
3	Sum^{Cas}	0.16	0.15	0.13	0.14	0.18	0.18	0.20	0.19	0.19	0.11	0.14	0.14	0.18	0.18	0.20	0.21	0.22	0.21	0.13	0.18	0.18	0.18	0.14	0.26
	Sum^{Diff}	0.15	0.15	0.13	0.14	0.17	0.18	0.20	0.19	0.19	0.10	0.12	0.14	0.18	0.18	0.20	0.19	0.22	0.21	0.13	0.19	0.18	0.18	0.16	0.27
	M	0.16	0.15	0.13	0.14	0.18	0.18	0.20	0.19	0.18	0.11	0.14	0.14	0.18	0.18	0.20	0.21	0.22	0.21	0.13	0.19	0.18	0.18	0.14	0.26
7	Sum^{Cas}	0.17	0.15	0.14	0.14	0.15	0.18	0.19	0.18	0.18	0.17	0.16	0.18	0.19	0.22	0.23	0.22	0.23	0.23	0.23	0.25	0.17	0.21	0.07	0.13
	Sum^{Diff}	0.17	0.14	0.14	0.13	0.15	0.16	0.19	0.17	0.18	0.17	0.16	0.16	0.18	0.21	0.22	0.22	0.21	0.21	0.22	0.23	0.16	0.21	0.07	0.14
	M	0.17	0.15	0.14	0.14	0.15	0.18	0.19	0.18	0.18	0.17	0.16	0.18	0.19	0.22	0.23	0.22	0.23	0.23	0.23	0.25	0.17	0.21	0.07	0.14
11	Sum^{Cas}	0.16	0.15	0.16	0.14	0.16	0.13	0.16	0.19	0.19	0.18	0.18	0.19	0.19	0.19	0.21	0.19	0.21	0.23	0.22	0.23	0.23	0.23	0.05	0.10
	Sum^{Diff}	0.15	0.15	0.15	0.14	0.15	0.13	0.15	0.18	0.19	0.18	0.18	0.20	0.19	0.18	0.20	0.19	0.21	0.21	0.22	0.23	0.22	0.21	0.05	0.13
	M	0.16	0.15	0.16	0.14	0.16	0.13	0.16	0.19	0.19	0.18	0.18	0.19	0.19	0.19	0.21	0.19	0.21	0.23	0.22	0.23	0.23	0.23	0.05	0.09
Match																									
3	Sum^{Cas}	0.17	0.14	0.16	0.14	0.17	0.22	0.15	0.16	0.14	0.16	0.18	0.19	0.16	0.14	0.14	0.25	0.21	0.24	0.17	0.21	0.21	0.19	0.16	0.42
	Sum^{Diff}	0.17	0.14	0.15	0.14	0.17	0.22	0.16	0.16	0.14	0.16	0.19	0.19	0.16	0.15	0.14	0.26	0.22	0.25	0.16	0.18	0.20	0.20	0.17	0.44
	M	0.17	0.14	0.16	0.14	0.17	0.22	0.15	0.16	0.14	0.16	0.18	0.19	0.16	0.14	0.14	0.25	0.21	0.24	0.17	0.20	0.21	0.19	0.16	0.42
7	Sum^{Cas}	0.16	0.15	0.18	0.19	0.13	0.14	0.20	0.18	0.17	0.16	0.14	0.15	0.15	0.26	0.24	0.32	0.32	0.32	0.32	0.29	0.17	0.16	0.12	0.39
	Sum^{Diff}	0.16	0.15	0.18	0.19	0.13	0.13	0.20	0.17	0.16	0.15	0.14	0.16	0.14	0.22	0.24	0.28	0.29	0.30	0.30	0.28	0.16	0.16	0.12	0.41
	M	0.16	0.15	0.18	0.19	0.13	0.14	0.20	0.18	0.17	0.16	0.14	0.15	0.15	0.26	0.24	0.32	0.32	0.32	0.32	0.29	0.17	0.16	0.12	0.39
11	Sum^{Cas}	0.20	0.16	0.17	0.18	0.18	0.20	0.15	0.16	0.17	0.17	0.13	0.29	0.28	0.29	0.30	0.29	0.31	0.31	0.26	0.25	0.29	0.26	0.13	0.36
	Sum^{Diff}	0.19	0.16	0.17	0.17	0.18	0.20	0.15	0.16	0.17	0.16	0.13	0.23	0.23	0.27	0.26	0.27	0.29	0.29	0.25	0.24	0.28	0.22	0.11	0.34
	M	0.20	0.16	0.17	0.18	0.18	0.20	0.15	0.16	0.17	0.17	0.13	0.29	0.28	0.29	0.29	0.29	0.31	0.31	0.26	0.25	0.29	0.26	0.13	0.36

We assessed the impact of removing rare SNPs and SNPs that were in relatively high LD with each other, as described in Section 3.2.10. Applying the 2% MAF criterion, 19 SNPs were extracted, leaving 295, a 6.1% reduction. Of the 22 selected SNPs within and surrounding the BRCA1 gene in the unpruned data sets, SNP 9 (rs8176225) was removed due to its MAF of 0.001, resulting in 21 referent SNPs for analysis, and the power results are presented in Table 3.7 for the $\log_{10}$(CHSS) reference marker score and summary statistics. The effect of removing rare SNPs from the data sets was substantial, power decreased by 0.23 and 0.36 for the $\text{Sum}^{\text{Cas}}/M$ and Sum^{Diff} summary statistics, as compared to the unpruned analysis in Table 3.3, resulting in the empirically adjusted powers of 0.68 and 0.45 for $\text{Sum}^{\text{Cas}}/M$ and Sum^{Diff}, respectively, which was previously 0.91 and 0.81.

TABLE 3.7: Power of the $\log_{10}$(CHSS) reference marker score and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M after the 2% MAF filter was imposed on the entire set of 314 SNPs in the simulated data sets. A total of 19 SNPs were excluded, reducing the number of SNPs to 295. Of the 22 selected SNPs within and surrounding the BRCA1 gene in the unpruned data sets, SNP 9 (rs8176225) was removed due to its MAF of 0.001, resulting in 21 referent SNPs for analysis. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

Summary	Marker Positions																					
Statistic	1	2	3	4	5	6	7	8	10	11	12	13	14	15	16	17	18	19	20	21	22	Bonf.	Emp.
Sum^{Cas}	0.47	0.45	0.48	0.67	0.64	0.66	0.65	0.61	0.64	0.59	0.71	0.70	0.71	0.72	0.68	0.70	0.72	0.71	0.74	0.72	0.65	0.45	0.68
Sum^{Diff}	0.29	0.29	0.30	0.37	0.37	0.37	0.38	0.37	0.37	0.34	0.41	0.41	0.41	0.41	0.38	0.40	0.43	0.42	0.44	0.42	0.39	0.24	0.45
M	0.47	0.45	0.48	0.67	0.64	0.66	0.65	0.61	0.64	0.59	0.71	0.70	0.71	0.72	0.68	0.70	0.72	0.71	0.74	0.72	0.65	0.45	0.68

TABLE 3.8: Power of the $\log_{10}$(CHSS) reference marker score and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M after the 1% LD filter was imposed on the entire set of 314 SNPs in the simulated data sets. A total of 277 SNPs were excluded, reducing the number of SNPs to 37. Of the 22 selected SNPs within and surrounding the BRCA1 gene in the unpruned data sets, 20 were removed resulting in only 2 referent SNPs left for analysis. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

Summary Statistic	Marker Positions			
	3	9	Bonf.	Emp.
Sum^{Cas}	0.21	0.24	0.18	0.21
Sum^{Diff}	0.15	0.18	0.14	0.17
M	0.21	0.25	0.18	0.21

We imposed the 1% LD filter on the full data set; 88.2% of the SNPs (N = 277) were pruned for which 37 of the SNPs remained in total and only 2 referent SNPs were left for analysis (Table 3.8). This had an even greater impact on power than the 2% MAF filter as the empirically adjusted powers of $\text{Sum}^{\text{Cas}}/M$ and Sum^{Diff} were 0.21 and 0.17, respectively.

In the last SNP exclusion scenario, we enforced both the 2% MAF and 1% LD criteria. A total of 285 SNPs were excluded (90.8% of the SNPs), leaving 29 SNPs in the entire data set and one SNP for analysis (Table 3.9). The same powers at SNP 3 were computed compared to when only the 1% LD filter was utilized (Table 3.8), 0.21 and 0.15 for $\text{Sum}^{\text{Cas}}/M$ and Sum^{Diff}, respectively.

The power of the allelic and haplotype χ^2 tests for fixed windows of sizes 3, 7, and 11 are shown in Table 3.10. Based on the Bonferroni and empirically adjusted powers, the allelic and haplotype χ^2 tests were about comparable, regardless of window size for the haplotype χ^2 test. However, in comparison to the reference marker scores ($\log_{10}$(CHSS), Length, and Count) using $\text{Sum}^{\text{Cas}}/M$ (Table 3.3) and both the binary and ratio threshold scores of $\log_{10}$(CHSS) (Table 3.5), both the unadjusted and adjusted

TABLE 3.9: Power of the $\log_{10}$(CHSS) reference marker score and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M after both the 2% MAF and 1% LD filters were imposed on the entire set of 314 SNPs in the simulated data sets. A total of 285 SNPs were excluded, reducing the number of SNPs to 29. Of the 22 selected SNPs within and surrounding the BRCA1 gene in the unpruned data sets, 21 were removed resulting in only 1 referent SNP left for analysis. One founder mutation carried throughout 100 generations was simulated.

Summary Statistic	Marker Position 3
Sum^{Cas}	0.21
Sum^{Diff}	0.15
M	0.21

powers of the allelic and haplotype χ^2 tests were considerably lower. On the other hand, the window based scores (Table 3.6) overall had lower powers than the allelic and haplotype χ^2 tests.

The power of our novel, quick, and efficient $R \times 2$ clustering algorithm and grouping techniques (No Regrouping, Regrouping, and Small Cluster Row) is presented in Table 3.11. For lower threshold values of 75%, 90%, and 99%, the powers calculated at each SNP and after multiple testing adjustments were unsatisfactory, ranging from an empirically adjusted power of 0.41 (90% threshold and Small Cluster Row/Keep) to 0.66 (99% threshold and Regrouping/Delete). In this threshold range from 75% to 99%, the Regrouping technique was consistently higher than the other 2 approaches (No Regrouping and Small Cluster Row) and as a matter of fact, its power remained stable at about 0.65, whereas the powers of the other 2 approaches steadily climbed. However, at the 99.5% threshold, all of the grouping techniques' powers jumped to above 0.80. At the higher 2 thresholds of 99.5% and 99.9%, the No Regrouping approach was the most powerful and not the Regrouping method that was the most powerful before. The empirically adjusted power of No Regrouping/Keep was 0.93 at the 99.5% threshold and then was 0.99 at the most stringent 99.9% threshold. In these higher thresholds, keeping

the "other" group of haplotypes in the $R \times 2$ tables seemed to be more powerful than deleting this group of the analysis entirely. On the contrary, for the prior thresholds from 75% to 99%, the Keep and Delete methods were similar.

In comparison to the permutation based approach of $\text{Sum}^{\text{Case}}/M$ using the $\log_{10}$(CHSS) reference marker (Table 3.3) and threshold scores (Table 3.5), the $R \times 2$ clustering algorithm at the highest 99.9% threshold performed better than the reference marker and ratio scores and was comparable to the binary scores. We note that the powers of the $R \times 2$ clustering algorithm reached above 0.90 at the 99.5% threshold though for the binary scores, higher powers were attained at the lower 99% threshold.

Finally, constructing 2×2 tables with the clustering algorithm did not prove to be a beneficial or competing approach (Table 3.12). The greatest power calculated was 0.74 at the 99.9% threshold and for the 1-sided test. At the lower thresholds from 75% to 99.5%, the empirical powers ranged from 0 to 0.21.

TABLE 3.10: Power of the allelic and haplotype χ^2 tests for fixed windows of sizes 3, 7, and 11 at each of the 22 selected SNPs within and surrounding the BRCA1 gene. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

	Marker Positions																							
Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Bonf.	Emp.
Allelic Test	0.30	0.11	0.25	0.24	0.28	0.26	0.26	0.29	0.00	0.26	0.16	0.27	0.26	0.19	0.27	0.30	0.18	0.09	0.28	0.04	0.27	0.25	0.36	0.55
Haplotype χ^2 Test																								
3	0.41	0.29	0.35	0.33	0.19	0.27	0.27	0.27	0.29	0.29	0.29	0.29	0.22	0.23	0.27	0.39	0.37	0.36	0.25	0.22	0.23	0.23	0.37	0.58
7	0.48	0.46	0.46	0.41	0.41	0.37	0.29	0.35	0.35	0.35	0.36	0.31	0.36	0.42	0.41	0.41	0.42	0.42	0.41	0.35	0.31	0.29	0.34	0.53
11	0.57	0.45	0.45	0.44	0.44	0.47	0.45	0.40	0.33	0.35	0.35	0.44	0.44	0.44	0.42	0.42	0.43	0.42	0.38	0.39	0.40	0.42	0.45	0.58

TABLE 3.11: Power of the $R \times 2$ clustering algorithm and grouping technique (No Regrouping, Regrouping, and Small Cluster Row) at each of the 22 selected SNPs within and surrounding the BRCA1 gene and for a range of thresholds (75%, 90%, 95%, 99%, 99.5%, and 99.9%) based on the $\log_{10}$(CHSS) reference marker score. The "other" group of haplotypes was both kept in and removed from the tables (Keep and Delete, respectively). Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

Threshold		Marker Positions																						
(%)	Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Bonf.	Emp.
75	No Regrouping																								
	Keep	0.27	0.10	0.21	0.20	0.21	0.21	0.21	0.27	0.14	0.21	0.30	0.23	0.24	0.23	0.23	0.25	0.18	0.05	0.26	0.23	0.22	0.20	0.33	0.49
	Delete	0.30	0.12	0.26	0.22	0.24	0.25	0.25	0.29	0.19	0.24	0.29	0.20	0.20	0.21	0.21	0.28	0.22	0.06	0.27	0.26	0.27	0.27	0.36	0.51
	Regrouping																								
	Keep	0.30	0.11	0.25	0.24	0.23	0.20	0.21	0.27	0.14	0.22	0.27	0.19	0.19	0.21	0.25	0.26	0.20	0.09	0.28	0.21	0.26	0.25	0.36	0.64
	Delete	0.30	0.11	0.25	0.24	0.23	0.24	0.24	0.27	0.19	0.22	0.29	0.21	0.19	0.21	0.25	0.26	0.20	0.09	0.28	0.25	0.26	0.25	0.36	0.65
	Small Cluster Row																								
	Keep	0.26	0.10	0.19	0.19	0.17	0.15	0.15	0.26	0.11	0.15	0.27	0.18	0.20	0.21	0.21	0.22	0.18	0.05	0.22	0.20	0.20	0.22	0.28	0.44
	Delete	0.29	0.12	0.25	0.21	0.19	0.18	0.18	0.26	0.14	0.20	0.28	0.21	0.21	0.22	0.22	0.24	0.19	0.05	0.26	0.22	0.22	0.23	0.35	0.46
90	No Regrouping																								
	Keep	0.29	0.13	0.25	0.21	0.22	0.22	0.22	0.27	0.16	0.19	0.28	0.19	0.19	0.19	0.19	0.26	0.21	0.06	0.20	0.21	0.20	0.20	0.31	0.49
	Delete	0.29	0.14	0.28	0.21	0.27	0.25	0.25	0.31	0.22	0.27	0.28	0.27	0.27	0.27	0.27	0.30	0.25	0.07	0.28	0.29	0.28	0.28	0.37	0.53
	Regrouping																								
	Keep	0.30	0.11	0.25	0.23	0.26	0.27	0.27	0.27	0.17	0.23	0.28	0.23	0.22	0.21	0.25	0.31	0.24	0.10	0.25	0.21	0.26	0.25	0.37	0.65
	Delete	0.30	0.11	0.25	0.23	0.26	0.27	0.27	0.27	0.21	0.24	0.28	0.24	0.23	0.21	0.25	0.30	0.24	0.10	0.25	0.26	0.26	0.25	0.37	0.65
	Small Cluster Row																								
	Keep	0.21	0.11	0.19	0.17	0.22	0.20	0.20	0.26	0.14	0.16	0.22	0.16	0.19	0.19	0.19	0.23	0.21	0.07	0.19	0.20	0.20	0.18	0.25	0.41
	Delete	0.28	0.11	0.25	0.16	0.21	0.21	0.21	0.29	0.15	0.19	0.28	0.19	0.20	0.20	0.20	0.27	0.24	0.07	0.23	0.22	0.23	0.22	0.30	0.46
95	No Regrouping																								
	Keep	0.29	0.12	0.23	0.23	0.24	0.24	0.24	0.27	0.25	0.25	0.32	0.26	0.26	0.26	0.26	0.32	0.25	0.17	0.27	0.26	0.27	0.28	0.32	0.51
	Delete	0.31	0.15	0.26	0.23	0.25	0.26	0.26	0.32	0.29	0.29	0.33	0.29	0.29	0.29	0.29	0.32	0.30	0.15	0.30	0.30	0.30	0.30	0.38	0.52
	Regrouping																								

(*continued*)

Threshold (%)	Method	Marker Positions 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Bonf.	Emp.
	Keep	0.30	0.11	0.26	0.25	0.25	0.26	0.26	0.26	0.25	0.26	0.30	0.25	0.24	0.27	0.27	0.31	0.29	0.15	0.28	0.22	0.29	0.26	0.37	0.64
	Delete	0.30	0.11	0.26	0.25	0.25	0.26	0.26	0.26	0.26	0.27	0.30	0.26	0.25	0.27	0.27	0.30	0.29	0.16	0.28	0.30	0.29	0.26	0.37	0.65
	Small Cluster Row																								
	Keep	0.25	0.14	0.24	0.25	0.25	0.25	0.25	0.29	0.27	0.27	0.32	0.28	0.29	0.29	0.29	0.30	0.28	0.17	0.28	0.28	0.27	0.27	0.28	0.45
	Delete	0.27	0.10	0.22	0.23	0.26	0.27	0.27	0.32	0.29	0.29	0.31	0.29	0.29	0.29	0.29	0.34	0.30	0.15	0.27	0.27	0.27	0.28	0.32	0.47
99	No Regrouping																								
	Keep	0.47	0.37	0.46	0.45	0.45	0.45	0.45	0.47	0.46	0.46	0.49	0.46	0.46	0.46	0.46	0.44	0.52	0.47	0.48	0.49	0.49	0.49	0.57	0.60
	Delete	0.44	0.36	0.44	0.39	0.39	0.39	0.39	0.41	0.40	0.40	0.45	0.38	0.37	0.37	0.37	0.36	0.45	0.46	0.49	0.50	0.50	0.50	0.55	0.62
	Regrouping																								
	Keep	0.38	0.37	0.36	0.32	0.33	0.32	0.32	0.35	0.34	0.33	0.38	0.31	0.32	0.32	0.30	0.31	0.44	0.38	0.38	0.41	0.38	0.37	0.39	0.61
	Delete	0.39	0.36	0.37	0.33	0.32	0.33	0.33	0.36	0.37	0.37	0.38	0.32	0.33	0.33	0.33	0.32	0.48	0.40	0.39	0.44	0.41	0.41	0.38	0.66
	Small Cluster Row																								
	Keep	0.42	0.37	0.42	0.46	0.45	0.45	0.45	0.46	0.46	0.46	0.49	0.46	0.46	0.46	0.46	0.47	0.50	0.48	0.48	0.48	0.48	0.48	0.54	0.63
	Delete	0.43	0.36	0.42	0.41	0.40	0.40	0.40	0.40	0.40	0.40	0.43	0.40	0.40	0.40	0.40	0.39	0.50	0.46	0.46	0.46	0.46	0.46	0.53	0.59
99.5	No Regrouping																								
	Keep	0.91	0.89	0.89	0.88	0.88	0.88	0.88	0.89	0.89	0.89	0.88	0.89	0.89	0.89	0.89	0.89	0.91	0.90	0.90	0.90	0.90	0.90	0.85	0.93
	Delete	0.85	0.83	0.83	0.84	0.84	0.84	0.84	0.83	0.83	0.83	0.83	0.83	0.82	0.82	0.82	0.82	0.85	0.84	0.84	0.84	0.84	0.84	0.72	0.87
	Regrouping																								
	Keep	0.73	0.71	0.71	0.72	0.72	0.71	0.71	0.74	0.72	0.73	0.76	0.74	0.74	0.74	0.74	0.72	0.75	0.75	0.74	0.76	0.73	0.73	0.69	0.83
	Delete	0.72	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.70	0.71	0.72	0.72	0.72	0.72	0.72	0.71	0.76	0.72	0.73	0.75	0.74	0.74	0.73	0.84
	Small Cluster Row																								
	Keep	0.89	0.87	0.86	0.88	0.88	0.88	0.88	0.90	0.89	0.89	0.89	0.88	0.88	0.88	0.88	0.88	0.91	0.90	0.90	0.90	0.90	0.90	0.86	0.91
	Delete	0.89	0.89	0.89	0.88	0.88	0.88	0.88	0.89	0.89	0.89	0.89	0.88	0.88	0.88	0.88	0.88	0.93	0.90	0.90	0.90	0.90	0.90	0.85	0.91
99.9	No Regrouping																								
	Keep	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.99
	Delete	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.94	0.94	0.94	0.94	0.94	0.94	0.90	0.93
	Regrouping																								
	Keep	0.70	0.71	0.72	0.73	0.67	0.67	0.67	0.65	0.69	0.69	0.73	0.71	0.70	0.72	0.72	0.73	0.75	0.76	0.72	0.79	0.76	0.75	0.70	0.91
	Delete	0.70	0.67	0.68	0.71	0.68	0.69	0.69	0.65	0.68	0.68	0.72	0.70	0.70	0.69	0.71	0.69	0.75	0.72	0.73	0.75	0.74	0.73	0.68	0.88
	Small Cluster Row																								
	Keep	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.92	0.98
	Delete	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.94	0.94	0.94	0.72	0.86

TABLE 3.12: Power of the 2×2 clustering algorithm at each of the 22 selected SNPs within and surrounding the BRCA1 gene and for a range of thresholds (75%, 90%, 95%, 99%, 99.5%, and 99.9%) based on the $\log_{10}$(CHSS) reference marker score. Both 1- and 2-sided tests were performed. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

Threshold		Marker Positions																						
(%)	Hypothesis	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Bonf.	Emp.
75	2-Sided	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.01	0.00	0.00
	1-Sided	0.05	0.03	0.03	0.03	0.05	0.02	0.02	0.02	0.02	0.03	0.01	0.05	0.04	0.04	0.04	0.04	0.02	0.03	0.03	0.04	0.01	0.01	0.00	0.00
90	2-Sided	0.04	0.04	0.03	0.04	0.03	0.02	0.02	0.02	0.01	0.02	0.02	0.03	0.01	0.02	0.02	0.02	0.02	0.01	0.01	0.00	0.01	0.01	0.00	0.01
	1-Sided	0.08	0.07	0.08	0.10	0.10	0.09	0.09	0.11	0.10	0.09	0.08	0.08	0.10	0.10	0.10	0.10	0.08	0.08	0.09	0.05	0.06	0.06	0.01	0.04
95	2-Sided	0.09	0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.06	0.06	0.07	0.07	0.08	0.08	0.08	0.08	0.08	0.08	0.09	0.10	0.09	0.09	0.02	0.09
	1-Sided	0.14	0.11	0.08	0.19	0.17	0.15	0.15	0.15	0.14	0.14	0.14	0.14	0.13	0.13	0.13	0.14	0.12	0.13	0.14	0.14	0.14	0.16	0.03	0.15
99	2-Sided	0.11	0.13	0.12	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.11	0.14	0.15	0.15	0.15	0.14	0.12	0.14	0.14	0.15	0.15	0.15	0.04	0.13
	1-Sided	0.19	0.18	0.17	0.26	0.26	0.25	0.25	0.24	0.24	0.24	0.24	0.25	0.24	0.24	0.24	0.23	0.22	0.21	0.21	0.21	0.21	0.21	0.04	0.10
99.5	2-Sided	0.25	0.27	0.27	0.29	0.29	0.29	0.29	0.27	0.29	0.29	0.28	0.28	0.28	0.28	0.28	0.27	0.26	0.24	0.25	0.25	0.25	0.25	0.11	0.16
	1-Sided	0.37	0.39	0.38	0.38	0.38	0.38	0.38	0.36	0.37	0.37	0.37	0.38	0.37	0.37	0.37	0.37	0.37	0.34	0.33	0.33	0.33	0.33	0.14	0.21
99.9	2-Sided	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.73	0.73	0.73	0.73	0.73	0.75	0.75	0.75	0.75	0.75	0.75	0.42	0.70
	1-Sided	0.82	0.82	0.82	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.48	0.74

The same analytical approaches that were carried out for BRCA1 were also conducted for PHB and are included in Tables 3.13 through 3.22. Table 3.13 presents the power of the $\log_{10}$(CHSS), Length, and Count reference marker scores and accompanying summary statistics for the PHB simulated data sets. Overall, the powers were greater in the PHB than the BRCA1 simulated data sets. The majority of the empirically adjusted powers were above 0.87. Conversely, the Length reference marker score performed the best as its empirically adjusted powers for $\text{Sum}^{\text{Cas}}/M$ almost reached 1.00, whereas $\log_{10}$(CHSS) had the highest powers for BRCA1. Also, the powers of the Count reference marker score were considerably better at 0.90 (empirically adjusted) for $\text{Sum}^{\text{Cas}}/M$. The empirically adjusted powers of the $\log_{10}$(CHSS) reference marker score improved slightly by 0.03 ($\text{Sum}^{\text{Cas}}/M = 0.94$ and $\text{Sum}^{\text{Diff}} = 0.87$). Similar to our observation with BRCA1, the permutation based Sum^{Diff} was consistently not as powerful as Sum^{Cas} and M. For the data sets that incorporated 5 independent founder mutations carried throughout 100 generations (Table 3.14), the empirically adjusted powers of the $\log_{10}$(CHSS) and Length reference marker scores for $\text{Sum}^{\text{Cas}}/M$ were fairly similar. This was not the case with BRCA1 as the same patterns of results were seen with the 5 mutation models as with the 1 mutation models.

TABLE 3.13: Power of the $\log_{10}$(CHSS), Length, and Count reference marker scores and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M at each of the 16 selected SNPs within and surrounding the PHB gene. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

	Summary	Marker Positions																	
Method	Statistic	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Bonf.	Emp.
$\log_{10}$(CHSS)	Sum^{Cas}	0.89	0.88	0.88	0.94	0.93	0.92	0.97	0.93	0.92	0.96	0.94	0.93	0.96	0.91	0.92	0.93	0.86	0.94
	Sum^{Diff}	0.70	0.71	0.71	0.86	0.86	0.86	0.89	0.88	0.86	0.86	0.87	0.84	0.86	0.76	0.75	0.76	0.71	0.87
	M	0.89	0.88	0.88	0.94	0.93	0.92	0.97	0.93	0.92	0.96	0.94	0.93	0.96	0.91	0.93	0.93	0.86	0.94
Length	Sum^{Cas}	0.80	0.80	0.80	0.94	0.94	0.93	0.96	0.93	0.94	0.98	0.98	0.98	1.00	0.90	0.93	0.90	0.89	0.99
	Sum^{Diff}	0.67	0.68	0.67	0.86	0.86	0.86	0.89	0.87	0.86	0.85	0.87	0.85	0.87	0.77	0.76	0.80	0.72	0.87
	M	0.80	0.80	0.80	0.94	0.94	0.93	0.96	0.94	0.94	0.98	0.98	0.98	1.00	0.90	0.93	0.90	0.89	0.99
Count	Sum^{Cas}	0.79	0.80	0.80	0.94	0.93	0.93	0.96	0.93	0.92	0.95	0.95	0.92	0.94	0.82	0.87	0.87	0.86	0.90
	Sum^{Diff}	0.66	0.67	0.67	0.78	0.84	0.82	0.88	0.84	0.82	0.84	0.84	0.80	0.80	0.70	0.70	0.64	0.70	0.79
	M	0.79	0.80	0.80	0.94	0.93	0.93	0.96	0.93	0.92	0.95	0.95	0.92	0.94	0.82	0.87	0.87	0.86	0.90

TABLE 3.14: Power of the $\log_{10}$(CHSS), Length, and Count reference marker scores and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M at each of the 16 selected SNPs within and surrounding the PHB gene. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. Five independent founder mutations carried throughout 100 generations were simulated.

	Summary	Marker Positions																	
Method	Statistic	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Bonf.	Emp.
$\log_{10}$(CHSS)	Sum^{Cas}	0.87	0.90	0.90	0.96	0.97	0.99	0.98	0.97	0.97	0.98	0.99	0.97	0.99	0.96	0.93	0.96	0.96	0.99
	Sum^{Diff}	0.63	0.64	0.64	0.84	0.86	0.82	0.92	0.89	0.80	0.89	0.89	0.85	0.86	0.71	0.76	0.76	0.80	0.86
	M	0.87	0.90	0.90	0.96	0.97	0.99	0.98	0.97	0.97	0.98	0.99	0.97	0.99	0.96	0.93	0.96	0.96	0.99
Length	Sum^{Cas}	0.78	0.80	0.81	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.98	0.98	0.99	0.96	0.93	0.94	0.93	0.99
	Sum^{Diff}	0.53	0.58	0.58	0.86	0.83	0.83	0.86	0.85	0.87	0.91	0.89	0.88	0.88	0.76	0.76	0.78	0.73	0.90
	M	0.78	0.80	0.81	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.98	0.98	0.99	0.96	0.93	0.94	0.93	0.99
Count	Sum^{Cas}	0.75	0.82	0.82	0.93	0.98	0.95	0.99	0.98	0.97	0.97	0.96	0.95	0.96	0.88	0.88	0.86	0.92	0.93
	Sum^{Diff}	0.52	0.58	0.58	0.79	0.84	0.80	0.86	0.83	0.78	0.90	0.88	0.83	0.84	0.56	0.64	0.60	0.72	0.78
	M	0.75	0.82	0.82	0.93	0.98	0.95	0.99	0.98	0.97	0.97	0.96	0.95	0.96	0.88	0.88	0.86	0.92	0.93

Table 3.15 shows the power of the binary and ratio threshold scores based on the $\log_{10}$(CHSS) reference marker scores. As with BRCA1 (Table 3.5), the binary scores resulted in greater powers than the ratio scores for the more stringent thresholds in the PHB simulated data sets. However, for PHB the maximal powers of the binary scores occurred as of the 99.5% threshold in contrast to BRCA1 in which powers peaked earlier at the 99% threshold.

Table 3.16 contains the results of the window based scores, analogous to the BRCA1 results in Table 3.6. All of the observations described previously for BRCA1 also held for the PHB analysis, with the exception that the empirically adjusted powers of the Length and Match scores were almost equivalent for the smaller window sizes of 3 and 7 and these scores were also the most powerful. On the other hand, for the BRCA1 simulated data sets, the $\log_{10}$(CHSS) had the greatest power.

Table 3.17 exhibits the power of the $\log_{10}$(CHSS) reference marker score and accompanying summary statistics after the 2% MAF filter was imposed on the entire set of 486 SNPs in the PHB simulated data sets. In contrast to the BRCA1 results (Table 3.7), there was not a substantial decrease in power after removing rare variants for Sum^{Cas} and M as their power remained stable at 0.91 (without removing rare SNPs their power was 0.94). Though for Sum^{Diff} its power decreased considerably from 0.87 (Table 3.13) to 0.67.

In the second and third SNP exclusion scenario in which the 1% LD and 2% MAF/1% LD filters were applied (Tables 3.18 and 3.19), both the unadjusted and adjusted powers were approximately equivalent to those of the BRCA1 simulated and pruned data sets.

TABLE 3.15: Power of the $\log_{10}$(CHSS) reference marker threshold scores (binary and ratio) and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M at each of the 16 selected SNPs within and surrounding the PHB gene. The 75%, 90%, 95%, 99%, 99.5%, and 99.9% thresholds were considered. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

Threshold (%)	Summary Statistic	Marker Positions 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Bonf.	Emp.
Binary																			
75	Sum^{Cas}	0.29	0.30	0.30	0.33	0.37	0.37	0.34	0.31	0.39	0.35	0.34	0.31	0.31	0.22	0.19	0.19	0.30	0.39
	Sum^{Diff}	0.25	0.28	0.26	0.26	0.33	0.33	0.28	0.27	0.33	0.31	0.30	0.31	0.24	0.17	0.17	0.17	0.23	0.40
	M	0.29	0.30	0.30	0.33	0.37	0.37	0.34	0.31	0.39	0.35	0.34	0.31	0.31	0.22	0.19	0.19	0.30	0.39
90	Sum^{Cas}	0.31	0.33	0.33	0.39	0.43	0.42	0.39	0.39	0.34	0.32	0.34	0.33	0.36	0.30	0.31	0.30	0.20	0.36
	Sum^{Diff}	0.27	0.28	0.27	0.29	0.27	0.29	0.27	0.25	0.25	0.24	0.25	0.23	0.27	0.21	0.25	0.25	0.15	0.27
	M	0.31	0.33	0.33	0.39	0.43	0.42	0.39	0.39	0.34	0.32	0.34	0.33	0.36	0.30	0.31	0.30	0.20	0.36
95	Sum^{Cas}	0.54	0.54	0.54	0.54	0.56	0.56	0.56	0.55	0.57	0.58	0.59	0.59	0.59	0.58	0.55	0.54	0.33	0.45
	Sum^{Diff}	0.37	0.36	0.35	0.39	0.39	0.40	0.39	0.36	0.41	0.42	0.44	0.41	0.43	0.37	0.37	0.35	0.25	0.31
	M	0.55	0.54	0.55	0.54	0.56	0.56	0.56	0.55	0.57	0.58	0.59	0.59	0.59	0.59	0.55	0.54	0.34	0.45
99	Sum^{Cas}	0.96	0.96	0.96	0.97	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.94	0.97	0.99	0.90	0.92
	Sum^{Diff}	0.88	0.90	0.89	0.89	0.89	0.90	0.89	0.90	0.89	0.90	0.90	0.89	0.89	0.89	0.90	0.90	0.74	0.80
	M	0.96	0.96	0.96	0.97	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.94	0.97	0.99	0.90	0.91
99.5	Sum^{Cas}	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.99
	Sum^{Diff}	0.99	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.94	0.96
	M	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.99
99.9	Sum^{Cas}	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98
	Sum^{Diff}	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.98
	M	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98
Ratio																			
75	Sum^{Cas}	0.27	0.29	0.30	0.22	0.32	0.26	0.24	0.29	0.44	0.25	0.26	0.23	0.24	0.16	0.24	0.18	0.31	0.54
	Sum^{Diff}	0.24	0.30	0.29	0.21	0.30	0.19	0.21	0.22	0.35	0.25	0.23	0.20	0.22	0.13	0.22	0.16	0.30	0.52
	M	0.27	0.29	0.30	0.22	0.32	0.26	0.24	0.29	0.44	0.25	0.26	0.23	0.24	0.16	0.24	0.18	0.32	0.54
90	Sum^{Cas}	0.28	0.30	0.31	0.27	0.38	0.36	0.53	0.43	0.38	0.47	0.48	0.47	0.37	0.23	0.21	0.23	0.26	0.45
	Sum^{Diff}	0.26	0.27	0.29	0.22	0.28	0.29	0.34	0.31	0.31	0.36	0.34	0.32	0.27	0.20	0.19	0.21	0.20	0.42
	M	0.28	0.30	0.31	0.27	0.38	0.36	0.53	0.43	0.38	0.47	0.48	0.47	0.37	0.23	0.20	0.23	0.26	0.45
95	Sum^{Cas}	0.31	0.34	0.34	0.40	0.47	0.45	0.61	0.53	0.43	0.57	0.52	0.54	0.48	0.33	0.35	0.29	0.26	0.59
	Sum^{Diff}	0.25	0.28	0.28	0.28	0.30	0.28	0.42	0.36	0.32	0.39	0.38	0.38	0.33	0.28	0.24	0.26	0.21	0.45
	M	0.31	0.34	0.34	0.40	0.47	0.45	0.61	0.53	0.43	0.57	0.52	0.54	0.49	0.33	0.35	0.29	0.26	0.58
99	Sum^{Cas}	0.45	0.51	0.49	0.60	0.68	0.70	0.75	0.71	0.68	0.73	0.72	0.71	0.66	0.54	0.57	0.49	0.50	0.72
	Sum^{Diff}	0.36	0.37	0.36	0.41	0.47	0.43	0.55	0.50	0.45	0.54	0.49	0.50	0.50	0.37	0.41	0.37	0.33	0.56
	M	0.45	0.51	0.49	0.60	0.68	0.70	0.75	0.71	0.67	0.73	0.72	0.71	0.66	0.54	0.57	0.49	0.50	0.73
99.5	Sum^{Cas}	0.54	0.62	0.62	0.80	0.85	0.80	0.85	0.81	0.79	0.82	0.81	0.78	0.78	0.68	0.66	0.64	0.65	0.85
	Sum^{Diff}	0.43	0.48	0.45	0.52	0.59	0.57	0.65	0.59	0.57	0.65	0.61	0.60	0.57	0.50	0.49	0.41	0.42	0.65
	M	0.54	0.62	0.62	0.79	0.85	0.80	0.85	0.81	0.79	0.82	0.81	0.78	0.78	0.68	0.66	0.64	0.65	0.85
99.9	Sum^{Cas}	0.81	0.81	0.82	0.92	0.91	0.91	0.92	0.91	0.90	0.89	0.91	0.90	0.91	0.84	0.84	0.86	0.84	0.91
	Sum^{Diff}	0.59	0.68	0.67	0.82	0.84	0.81	0.87	0.84	0.81	0.83	0.85	0.81	0.80	0.71	0.70	0.67	0.66	0.87
	M	0.81	0.81	0.82	0.92	0.91	0.91	0.92	0.91	0.90	0.89	0.91	0.90	0.91	0.84	0.84	0.86	0.84	0.91

TABLE 3.16: Power of the $\log_{10}$(CHSS), Length, Count, and Match scores for fixed windows of sizes 3, 7, and 11 SNPs and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M at each of the 16 selected SNPs within and surrounding the PHB gene. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

Window Size	Summary Statistic	Marker Positions 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Bonf.	Emp.
$\log_{10}$(CHSS)																			
3	Sum^{Cas}	0.18	0.14	0.21	0.22	0.14	0.21	0.20	0.19	0.18	0.19	0.19	0.21	0.21	0.18	0.25	0.25	0.40	0.69
	Sum^{Diff}	0.10	0.12	0.15	0.13	0.14	0.14	0.14	0.10	0.10	0.13	0.20	0.21	0.16	0.13	0.19	0.16	0.36	0.58
	M	0.18	0.14	0.21	0.22	0.14	0.21	0.20	0.19	0.18	0.19	0.19	0.21	0.21	0.18	0.25	0.25	0.40	0.69
7	Sum^{Cas}	0.18	0.18	0.22	0.20	0.23	0.16	0.17	0.20	0.20	0.18	0.21	0.18	0.21	0.23	0.25	0.22	0.34	0.41
	Sum^{Diff}	0.14	0.13	0.15	0.13	0.11	0.05	0.11	0.14	0.14	0.17	0.12	0.10	0.13	0.15	0.17	0.18	0.29	0.37
	M	0.18	0.18	0.22	0.20	0.23	0.16	0.17	0.20	0.20	0.18	0.21	0.18	0.21	0.23	0.25	0.22	0.34	0.41
11	Sum^{Cas}	0.22	0.22	0.24	0.22	0.20	0.20	0.21	0.18	0.24	0.23	0.23	0.20	0.22	0.23	0.23	0.26	0.28	0.43
	Sum^{Diff}	0.18	0.13	0.10	0.08	0.04	0.05	0.09	0.13	0.11	0.09	0.11	0.08	0.15	0.15	0.16	0.14	0.15	0.30
	M	0.22	0.22	0.24	0.22	0.20	0.20	0.21	0.18	0.24	0.23	0.23	0.20	0.22	0.23	0.23	0.26	0.28	0.43
Length																			
3	Sum^{Cas}	0.38	0.09	0.13	0.22	0.25	0.17	0.18	0.47	0.52	0.28	0.21	0.21	0.17	0.21	0.20	0.18	0.43	0.77
	Sum^{Diff}	0.33	0.10	0.13	0.20	0.19	0.17	0.17	0.47	0.37	0.27	0.20	0.21	0.18	0.19	0.21	0.16	0.37	0.76
	M	0.38	0.09	0.13	0.22	0.25	0.17	0.18	0.47	0.52	0.28	0.21	0.21	0.17	0.21	0.20	0.18	0.43	0.77
7	Sum^{Cas}	0.29	0.30	0.28	0.22	0.24	0.27	0.51	0.55	0.24	0.25	0.22	0.25	0.25	0.25	0.27	0.27	0.33	0.68
	Sum^{Diff}	0.24	0.25	0.26	0.20	0.21	0.22	0.37	0.51	0.22	0.24	0.19	0.21	0.20	0.21	0.25	0.25	0.33	0.64
	M	0.29	0.30	0.28	0.22	0.24	0.27	0.51	0.55	0.24	0.25	0.22	0.25	0.25	0.25	0.27	0.27	0.33	0.68
11	Sum^{Cas}	0.32	0.31	0.32	0.34	0.36	0.34	0.38	0.32	0.26	0.25	0.27	0.30	0.30	0.29	0.29	0.33	0.25	0.26
	Sum^{Diff}	0.28	0.28	0.30	0.30	0.33	0.29	0.29	0.24	0.24	0.22	0.24	0.23	0.23	0.24	0.26	0.27	0.22	0.22
	M	0.32	0.31	0.32	0.34	0.36	0.34	0.39	0.32	0.26	0.25	0.27	0.30	0.30	0.29	0.29	0.33	0.25	0.25
Count																			
3	Sum^{Cas}	0.21	0.11	0.12	0.20	0.14	0.21	0.18	0.31	0.29	0.31	0.22	0.20	0.14	0.18	0.16	0.20	0.30	0.39
	Sum^{Diff}	0.15	0.11	0.11	0.20	0.12	0.19	0.18	0.28	0.26	0.28	0.22	0.18	0.13	0.17	0.17	0.20	0.29	0.35
	M	0.21	0.11	0.12	0.20	0.14	0.21	0.18	0.31	0.29	0.31	0.22	0.20	0.14	0.18	0.17	0.20	0.30	0.39
7	Sum^{Cas}	0.15	0.15	0.22	0.24	0.23	0.26	0.23	0.27	0.27	0.26	0.28	0.30	0.23	0.22	0.24	0.22	0.25	0.31
	Sum^{Diff}	0.13	0.12	0.19	0.24	0.20	0.16	0.20	0.25	0.25	0.26	0.22	0.21	0.19	0.20	0.22	0.20	0.21	0.28
	M	0.15	0.15	0.22	0.24	0.23	0.26	0.23	0.27	0.27	0.26	0.28	0.30	0.23	0.22	0.24	0.22	0.25	0.30
11	Sum^{Cas}	0.17	0.24	0.25	0.28	0.27	0.30	0.26	0.23	0.26	0.29	0.26	0.26	0.31	0.28	0.22	0.26	0.22	0.25
	Sum^{Diff}	0.17	0.20	0.21	0.19	0.22	0.23	0.25	0.19	0.23	0.24	0.22	0.21	0.24	0.23	0.21	0.24	0.19	0.22
	M	0.17	0.24	0.25	0.28	0.27	0.30	0.26	0.23	0.26	0.29	0.26	0.26	0.31	0.28	0.22	0.26	0.22	0.24
Match																			
3	Sum^{Cas}	0.29	0.18	0.16	0.33	0.25	0.25	0.28	0.42	0.39	0.48	0.18	0.21	0.17	0.24	0.22	0.25	0.34	0.73
	Sum^{Diff}	0.28	0.18	0.17	0.31	0.17	0.18	0.21	0.30	0.32	0.32	0.18	0.20	0.17	0.22	0.21	0.24	0.31	0.72
	M	0.29	0.18	0.16	0.33	0.25	0.25	0.28	0.42	0.39	0.48	0.18	0.21	0.17	0.24	0.22	0.25	0.34	0.72
7	Sum^{Cas}	0.43	0.43	0.36	0.27	0.24	0.24	0.23	0.26	0.30	0.30	0.31	0.39	0.33	0.36	0.40	0.38	0.40	0.69
	Sum^{Diff}	0.37	0.34	0.32	0.24	0.19	0.19	0.19	0.25	0.28	0.21	0.27	0.33	0.29	0.30	0.31	0.27	0.34	0.66
	M	0.43	0.43	0.36	0.27	0.24	0.24	0.23	0.26	0.30	0.30	0.31	0.39	0.33	0.36	0.40	0.38	0.40	0.69
11	Sum^{Cas}	0.50	0.43	0.49	0.41	0.44	0.25	0.29	0.30	0.33	0.35	0.35	0.45	0.45	0.43	0.43	0.37	0.41	0.65
	Sum^{Diff}	0.37	0.31	0.33	0.33	0.35	0.22	0.20	0.22	0.26	0.28	0.27	0.32	0.33	0.34	0.36	0.26	0.32	0.60
	M	0.50	0.42	0.49	0.41	0.44	0.25	0.29	0.30	0.33	0.34	0.35	0.45	0.45	0.43	0.42	0.36	0.41	0.65

TABLE 3.17: Power of the $\log_{10}$(CHSS) reference marker score and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M after the 2% MAF filter was imposed on the entire set of 486 SNPs in the simulated data sets. A total of 12 SNPs were excluded, reducing the number of SNPs to 474. Of the 16 selected SNPs within and surrounding the PHB gene in the unpruned data sets, SNP 11 (rs2277636) was removed due to its MAF of 0.009, resulting in 15 referent SNPs for analysis. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

Summary	Marker Positions																
Statistic	1	2	3	4	5	6	7	8	9	10	12	13	14	15	16	Bonf.	Emp.
Sum^{Cas}	0.86	0.82	0.81	0.87	0.89	0.81	0.90	0.92	0.90	0.88	0.89	0.88	0.79	0.80	0.77	0.79	0.91
Sum^{Diff}	0.63	0.62	0.62	0.59	0.66	0.64	0.63	0.68	0.75	0.68	0.70	0.68	0.61	0.59	0.58	0.55	0.67
M	0.86	0.82	0.81	0.87	0.89	0.81	0.90	0.92	0.90	0.88	0.89	0.88	0.79	0.80	0.78	0.79	0.91

The allelic and haplotype χ^2 tests were also carried out for PHB (Table 3.20). The allelic test performed similarly for both gene regions; the empirically adjusted powers were 0.59 and 0.55 for PHB and BRCA1 (Table 3.10), respectively. However, for the haplotype χ^2 test computed using a window consisting of 3 adjacent SNPs, the empirically adjusted powers were almost the same for PHB and BRCA1 (0.62 and 0.58, respectively) and then for the PHB simulated data sets the empirically adjusted powers steadily climbed from 0.69 (window size 7) to 0.76 (window size 11), while on the contrary for the BRCA1 simulated data sets the empirically adjusted powers remained under 0.60.

The $R \times 2$ clustering algorithm was also applied to the PHB simulated data sets and the results are in Table 3.21. Similar to the BRCA1 analysis, the powers using the range of thresholds from 75% to 99% revealed inadequate powers as the maximum empirically adjusted power computed was 0.76 for No Regrouping and Keep/Delete at the 99% threshold. The empirically adjusted powers then improved significantly as of the 99.5% threshold. The majority of the empirically adjusted powers were above 0.90, the maximum empirically adjusted power occurring at the 99.9% threshold for the No Regrouping/Keep approach. Similarly to BRCA1, the No Regrouping technique was amongst the most powerful for the highest 2 thresholds of 99.5% and 99.9%. However,

TABLE 3.18: Power of the $\log_{10}$(CHSS) reference marker score and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M after the 1% LD filter was imposed on the entire set of 486 SNPs in the simulated data sets. A total of 430 SNPs were excluded, reducing the number of SNPs to 56. Of the 16 selected SNPs within and surrounding the PHB gene in the unpruned data sets, 14 were removed resulting in only 2 referent SNPs left for analysis. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

Summary Statistic	Marker Positions			
	11	15	Bonf.	Emp.
Sum^{Cas}	0.17	0.20	0.17	0.20
Sum^{Diff}	0.11	0.13	0.11	0.16
M	0.17	0.20	0.17	0.21

TABLE 3.19: Power of the $\log_{10}$(CHSS) reference marker score and accompanying summary statistics, Sum^{Cas}, Sum^{Diff}, and M after both the 2% MAF and 1% LD filters were imposed on the entire set of 486 SNPs in the simulated data sets. A total of 438 SNPs were excluded, reducing the number of SNPs to 48. Of the 16 selected SNPs within and surrounding the PHB gene in the unpruned data sets, 15 were removed resulting in only 1 referent SNP left for analysis. One founder mutation carried throughout 100 generations was simulated.

Summary Statistic	Marker Position
	15
Sum^{Cas}	0.23
Sum^{Diff}	0.16
M	0.23

TABLE 3.20: Power of the allelic and haplotype χ^2 tests for fixed windows of sizes 3, 7, and 11 at each of the 16 selected SNPs within and surrounding the PHB gene. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

	Marker Positions																	
Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Bonf.	Emp.
Allelic Test	0.23	0.14	0.14	0.21	0.28	0.17	0.33	0.33	0.28	0.36	0.03	0.32	0.32	0.15	0.28	0.11	0.48	0.59
Haplotype χ^2 Test																		
3	0.34	0.27	0.30	0.47	0.38	0.42	0.45	0.49	0.53	0.44	0.31	0.20	0.03	0.00	0.36	0.40	0.53	0.62
7	0.54	0.51	0.49	0.54	0.47	0.53	0.50	0.55	0.49	0.46	0.50	0.54	0.53	0.55	0.59	0.59	0.63	0.69
11	0.56	0.55	0.56	0.57	0.56	0.55	0.50	0.50	0.55	0.61	0.61	0.66	0.70	0.69	0.60	0.67	0.66	0.76

the Small Cluster Row had similar empirically adjusted powers as the No Regrouping technique, which we did not observe with the BRCA1 analysis. In addition, for the thresholds between 75% and 95%, there was not a clear distinction amongst the three competing grouping approaches, whereas for BRCA1 the Regrouping technique was consistently higher at these threshold levels. Lastly, keeping and deleting the "other" group of haplotypes in the $R \times 2$ tables had similar powers throughout the entire range of thresholds.

Finally, in Table 3.22 the powers of the 2×2 clustering algorithm are presented for the PHB simulated data sets. The empirically adjusted powers were quite similar to those of the BRCA1 analysis in Table 3.12, revealing poor power for this method in that none of the empirically adjusted powers reached 0.70, even for the higher 99.9% threshold.

TABLE 3.21: Power of the $R \times 2$ clustering algorithm and grouping technique (No Regrouping, Regrouping, and Small Cluster Row) at each of the 16 selected SNPs within and surrounding the PHB gene and for a range of thresholds (75%, 90%, 95%, 99%, 99.5%, and 99.9%) based on the $\log_{10}$(CHSS) reference marker score. The "other" group of haplotypes was both kept in and removed from the tables (Keep and Delete, respectively). Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

Threshold		Marker Positions																	
(%)	Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Bonf.	Emp.
75	No Regrouping																		
	Keep	0.23	0.14	0.14	0.21	0.28	0.17	0.32	0.32	0.28	0.34	0.03	0.30	0.31	0.15	0.28	0.11	0.48	0.58
	Delete	0.23	0.14	0.14	0.21	0.28	0.17	0.33	0.33	0.28	0.36	0.03	0.32	0.32	0.15	0.28	0.11	0.48	0.58
	Regrouping																		
	Keep	0.23	0.14	0.14	0.21	0.28	0.17	0.33	0.33	0.28	0.36	0.03	0.32	0.32	0.15	0.28	0.11	0.48	0.59
	Delete	0.23	0.14	0.14	0.21	0.28	0.17	0.33	0.33	0.28	0.36	0.03	0.32	0.32	0.15	0.28	0.11	0.48	0.59
	Small Cluster Row																		
	Keep	0.23	0.14	0.14	0.21	0.28	0.17	0.32	0.32	0.28	0.34	0.03	0.30	0.31	0.15	0.28	0.11	0.48	0.58
	Delete	0.23	0.14	0.14	0.21	0.28	0.17	0.33	0.33	0.28	0.36	0.03	0.31	0.32	0.15	0.28	0.11	0.48	0.59
90	No Regrouping																		
	Keep	0.19	0.14	0.14	0.22	0.26	0.15	0.26	0.26	0.26	0.25	0.04	0.28	0.23	0.15	0.26	0.11	0.44	0.53
	Delete	0.23	0.14	0.14	0.21	0.39	0.18	0.32	0.32	0.30	0.36	0.05	0.27	0.32	0.16	0.28	0.11	0.53	0.62
	Regrouping																		
	Keep	0.24	0.14	0.14	0.21	0.35	0.17	0.32	0.33	0.28	0.33	0.03	0.29	0.32	0.15	0.28	0.11	0.53	0.61
	Delete	0.24	0.14	0.14	0.21	0.35	0.17	0.32	0.32	0.29	0.33	0.04	0.30	0.32	0.15	0.28	0.11	0.53	0.61
	Small Cluster Row																		
	Keep	0.20	0.13	0.13	0.18	0.28	0.14	0.20	0.27	0.26	0.23	0.08	0.21	0.20	0.13	0.21	0.11	0.41	0.52
	Delete	0.19	0.14	0.14	0.17	0.35	0.18	0.28	0.27	0.25	0.28	0.04	0.25	0.22	0.15	0.25	0.12	0.50	0.59
95	No Regrouping																		
	Keep	0.28	0.18	0.18	0.27	0.43	0.33	0.38	0.41	0.44	0.35	0.16	0.33	0.28	0.17	0.26	0.11	0.54	0.57
	Delete	0.23	0.18	0.18	0.25	0.45	0.32	0.38	0.43	0.44	0.36	0.17	0.32	0.29	0.18	0.28	0.12	0.51	0.60
	Regrouping																		
	Keep	0.24	0.15	0.15	0.25	0.37	0.22	0.30	0.35	0.36	0.29	0.07	0.30	0.27	0.17	0.25	0.11	0.46	0.56
	Delete	0.25	0.16	0.15	0.26	0.40	0.26	0.35	0.40	0.38	0.32	0.11	0.31	0.28	0.18	0.25	0.11	0.48	0.59
	Small Cluster Row																		
	Keep	0.27	0.18	0.18	0.28	0.37	0.25	0.34	0.39	0.45	0.29	0.12	0.22	0.23	0.14	0.23	0.15	0.49	0.56
	Delete	0.25	0.18	0.18	0.24	0.43	0.28	0.34	0.39	0.43	0.35	0.09	0.26	0.25	0.16	0.24	0.14	0.49	0.59
99	No Regrouping																		
	Keep	0.58	0.59	0.59	0.67	0.73	0.71	0.77	0.77	0.79	0.77	0.70	0.71	0.72	0.57	0.59	0.52	0.71	0.76
	Delete	0.60	0.55	0.55	0.68	0.71	0.70	0.75	0.76	0.77	0.73	0.67	0.67	0.68	0.56	0.56	0.48	0.69	0.76
	Regrouping																		
	Keep	0.52	0.52	0.52	0.57	0.66	0.64	0.62	0.66	0.65	0.59	0.57	0.56	0.53	0.50	0.47	0.41	0.65	0.68
	Delete	0.51	0.48	0.48	0.56	0.67	0.64	0.65	0.67	0.64	0.57	0.57	0.57	0.55	0.52	0.46	0.43	0.66	0.70
	Small Cluster Row																		
	Keep	0.61	0.58	0.58	0.65	0.70	0.68	0.73	0.77	0.79	0.74	0.68	0.68	0.69	0.54	0.57	0.50	0.71	0.74
	Delete	0.60	0.54	0.54	0.64	0.72	0.68	0.70	0.73	0.75	0.71	0.66	0.66	0.68	0.53	0.55	0.50	0.70	0.74

(*continued*)

Threshold (%)	Method	Marker Positions 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Bonf.	Emp.
99.5	No Regrouping																		
	Keep	0.88	0.90	0.90	0.94	0.95	0.96	0.96	0.96	0.96	0.95	0.93	0.94	0.93	0.95	0.92	0.89	0.92	0.93
	Delete	0.87	0.87	0.87	0.93	0.94	0.95	0.94	0.94	0.94	0.93	0.91	0.91	0.90	0.93	0.91	0.86	0.85	0.90
	Regrouping																		
	Keep	0.70	0.68	0.68	0.75	0.77	0.78	0.76	0.80	0.82	0.81	0.80	0.75	0.75	0.78	0.76	0.73	0.80	0.84
	Delete	0.67	0.73	0.73	0.75	0.78	0.78	0.77	0.78	0.81	0.80	0.80	0.75	0.75	0.77	0.75	0.75	0.79	0.83
	Small Cluster Row																		
	Keep	0.89	0.89	0.89	0.95	0.94	0.95	0.95	0.96	0.96	0.95	0.93	0.93	0.93	0.93	0.92	0.88	0.90	0.93
	Delete	0.87	0.89	0.89	0.93	0.93	0.94	0.94	0.94	0.94	0.93	0.92	0.93	0.92	0.93	0.92	0.88	0.88	0.88
99.9	No Regrouping																		
	Keep	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.99
	Delete	0.98	0.98	0.98	0.98	0.97	0.98	0.98	0.98	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.95	0.96
	Regrouping																		
	Keep	0.78	0.83	0.83	0.83	0.84	0.84	0.80	0.78	0.87	0.82	0.83	0.80	0.79	0.78	0.72	0.76	0.86	0.90
	Delete	0.76	0.81	0.81	0.84	0.85	0.85	0.77	0.75	0.80	0.80	0.77	0.78	0.77	0.78	0.74	0.76	0.87	0.91
	Small Cluster Row																		
	Keep	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.97	0.98
	Delete	0.97	0.97	0.97	0.97	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.95	0.95	0.94	0.96	0.96	0.86	0.93

TABLE 3.22: Power of the 2×2 clustering algorithm at each of the 16 selected SNPs within and surrounding the PHB gene and for a range of thresholds (75%, 90%, 95%, 99%, 99.5%, and 99.9%) based on the $\log_{10}$(CHSS) reference marker score. Both 1- and 2-sided tests were performed. Bonferroni (Bonf.) and empirically (Emp.) adjusted powers are included, that accounted for the multiple testing across SNPs. One founder mutation carried throughout 100 generations was simulated.

Threshold		Marker Positions																	
(%)	Hypothesis	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Bonf.	Emp.
75	2-Sided	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1-Sided	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
90	2-Sided	0.06	0.05	0.05	0.08	0.04	0.04	0.05	0.08	0.10	0.06	0.09	0.07	0.04	0.04	0.01	0.01	0.02	0.03
	1-Sided	0.12	0.10	0.10	0.15	0.08	0.08	0.09	0.12	0.13	0.08	0.16	0.11	0.09	0.12	0.04	0.03	0.03	0.12
95	2-Sided	0.05	0.07	0.07	0.06	0.06	0.05	0.05	0.06	0.05	0.05	0.08	0.06	0.05	0.08	0.04	0.02	0.00	0.07
	1-Sided	0.11	0.11	0.11	0.16	0.15	0.15	0.10	0.13	0.11	0.14	0.13	0.09	0.12	0.09	0.10	0.12	0.03	0.14
99	2-Sided	0.10	0.12	0.12	0.10	0.13	0.11	0.14	0.15	0.14	0.14	0.14	0.14	0.14	0.16	0.14	0.20	0.04	0.13
	1-Sided	0.23	0.19	0.19	0.21	0.20	0.19	0.21	0.21	0.24	0.23	0.25	0.22	0.21	0.21	0.22	0.24	0.08	0.20
99.5	2-Sided	0.30	0.32	0.32	0.28	0.33	0.30	0.31	0.31	0.31	0.28	0.28	0.29	0.28	0.33	0.28	0.26	0.14	0.16
	1-Sided	0.42	0.41	0.41	0.41	0.40	0.41	0.40	0.41	0.37	0.39	0.39	0.40	0.40	0.43	0.39	0.42	0.16	0.31
99.9	2-Sided	0.68	0.69	0.69	0.75	0.73	0.73	0.75	0.76	0.76	0.77	0.77	0.78	0.77	0.77	0.78	0.76	0.54	0.65
	1-Sided	0.83	0.83	0.83	0.84	0.81	0.82	0.82	0.83	0.85	0.86	0.86	0.85	0.84	0.85	0.84	0.84	0.61	0.69

3.4 Discussion

We have demonstrated the utility of the $\log_{10}$(CHSS) reference marker score, recoding the $\log_{10}$(CHSS) as a threshold score, and the $R \times 2$ clustering algorithm. These approaches were powerful when analyzing data sets of unrelated cases and controls simulated under a rare and highly penetrant recessive mode-of-inheritance disease model for which the disease harboring haplotypes arose from founder populations. We hypothesized that generating haplotype sharing scores for all $\binom{4N}{2}$ pairs of haplotypes from these simulated unrelated cases and controls would result in a powerful approach since the genetic area flanking the founder mutation within the gene of interest would be highly conserved amongst the cases throughout the recombination events across 100 generations. The simulated rare, highly penetrant, and recessive disease model distributed the disease harboring haplotypes, that arose from the simulated founder pools, primarily amongst the cases. Therefore, the haplotype sharing measures were powered to detect excess sharing in the cases as compared to the controls who mainly carried the control haplotypes sampled from the general population.

In addition, we observed that the $\log_{10}$(CHSS) reference marker score outperformed the Length and Count in the BRCA1 analysis (Table 3.3), whereas all three scores were sufficiently powered in the PHB analysis (Table 3.13). BRCA1 is a more conserved region than the PHB gene. Therefore, we would expect that haplotype sharing amongst the pairs of case haplotypes would be somewhat comparable to the amount of sharing amongst the pairs of control haplotypes since the highly conserved region within BRCA1 would exhibit a similar haplotype architecture in both the case and control haplotypes. The result of this was that the Length and Count measures did not adequately distinguish excess sharing amongst the cases versus the controls. On the other hand, the PHB gene is much less conserved than BRCA1, thus the case haplotypes that originated from a single mutated founder haplotype would share many

more alleles than pairs of control haplotypes that likely did not resemble each other because they arose from a sample of haplotypes that had different haplotype signatures in this less conserved gene region. That said, we would expect the Length and Count to sufficiently distinguish excess haplotype sharing amongst the cases than the controls, which we observed as the Length and Count reference marker scores had robust powers in the PHB analysis, as well as $\log_{10}$(CHSS).

Furthermore, $\log_{10}$(CHSS) did not suffer a loss in power as the Length and Count did for the BRCA1 analysis, due to the rare SNPs contributing substantial weight to the computed $\log_{10}$(CHSS) scores in contrast to the Length and Count that were not designed to emphasize rare variants. This was clearly evident when we removed the rare SNPs by applying the 2% MAF criterion to the BRCA1 simulated data sets, causing the power of the $\log_{10}$(CHSS) to drop significantly as shown in Table 3.7.

We confirmed Lange and Boehnke's (2004) claim that Sum^{Diff} is not as powerful as Sum^{Cas}. They showed this in the parent-parent-affected offspring trio design whilst we have shown this in the unrelated case-control study design. Analogous to Lange and Boehnke's postulate, under the alternative hypothesis, the within group similarity for the groups of case and control haplotypes is high while the between group similarity is low. Thus, when subtracting off the sum of the control scores, the computed difference is small which does not reveal this grouping effect.

We found that the Mantel statistic (Mantel, 1967), M, advocated by Beckmann et al. (2005) had similar power to Sum^{Cas}. This was not surprising since by definition for the rare disease that we simulated with disease prevalence $K = 0.001$, the weights applied to the entire set of $\binom{4N}{2}$ pairs of haplotype scores were 0.998 (close to one) if both haplotypes were from cases, 10^{-6} (close to zero) if both haplotypes were from controls, and -0.000999 if the haplotypes were from discordant affected individuals (Section 3.2.4). Essentially, the approximate zero weight for the pairs of control haplo-

types removed their corresponding scores from the sum and the scores from discordant individuals were given very little weight, which minimally decreased the overall sum. Thus, the values of M were close to the values of Sum^{Cas}.

The binary threshold scores recoded from the $\log_{10}$(CHSS) reference marker scores was an incredibly powerful approach when the thresholds were set at a sufficiently high enough percentile. The binary score discretely separated the pairs of haplotypes that had relatively high scores from those that did not. These scores of greater magnitude were likely due to excess sharing amongst a small set of case haplotypes that comprised rare SNPs that matched. Therefore, upon shuffling the affection status labels in the permutation testing, the likelihood that all of these case haplotypes would have been assigned case labels was small relative to the assignment of some mixture of case/case, case/control, or control/control labels, which would not result in a summary statistic as high or greater than that observed.

Although the threshold ratio scores had excellent power at the highest 99.9% threshold, in general it did not perform as well as the binary scores. For scores that did not meet the given threshold, they were divided by the threshold in order to produce a distribution of ratios between zero and one. Scores that met or exceeded the threshold were simply assigned a value of one. In other words, scores were separated into groups of relatively high and low haplotype sharing, much like the binary score. However, upon permuting the affection status labels in order to assess statistical significance, the results suggest that the ratios would by chance, more often than not, inflate the permuted summary test statistic, thus weakening the power of the ratio score. In contrast, the binary score did not allow the smaller scores to augment the summary statistic since the smaller scores were assigned a value of zero.

Restricting the haplotype scoring to fixed windows of a specified number of adjacent SNPs proved to be an inadequate approach. For pairs of haplotypes that matched

for longer stretches of alleles than the window's coverage, the magnitude of the computed score was limited, therefore restricting the size of the score and subsequently the power of the summary statistic. In other words, the window based scores did not sufficiently distinguish between haplotype pairs that matched for great lengths and pairs that matched to a lesser degree. Although defining windows lends itself well to the theoretical tractability of the test statistics, we do not recommend the window based approach in haplotype sharing.

For the $\log_{10}$(CHSS) reference marker score, we showed that removing rare SNPs could adversely affect power. Though we did not demonstrate this, the Length and Count reference marker scores also suffered substantial losses in power regardless of the LD structure, physical size, and number of rare alleles in the gene region analyzed (i.e. losses were observed in both the BRCA1 and PHB analyses). Excluding the rare SNPs likely results in pairs of haplotypes that originally shared a great deal of alleles to not match as much, for which the altered case/case and control/control pairings would have the greatest effect on reducing power.

Pruning SNPs based on pairwise r^2 was also not desirable. The motivation for doing so would be to reduce the number of referent SNPs to analyze, thus alleviating the multiple testing burden. However, despite the multiple testing burden being minimal to none, the calculated power was extremely low.

We have proposed a fast and efficient algorithm that clusters similar haplotypes based on applying thresholds to the $\log_{10}$(CHSS) reference marker score. This method of constructing $R \times 2$ contingency tables was adequately powerful as of the 99.5% threshold and reached almost 1.00 for the 99.9% threshold. At these higher thresholds, the clustering algorithm was powered to detect discrete clusters of haplotypes in the simulated study subjects and differentiate between the specific sharing in the cases and controls. The motivation behind the clustering algorithm was similar to

the threshold scores that essentially separated the score data into groups of high and low sharing. The clustering algorithm further developed this idea of dichotomizing the data, whereby allowing the haplotype pairs with excess sharing to additionally be categorized into separate classes. In essence, more information was extracted from the score data by using the clustering algorithm than with the threshold score coupled with the summary statistic which required computationally intensive permutation testing to assess statistical significance. In contrast, the clustering algorithm was rapid since the test statistic of the $R \times 2$ table was Pearson's χ^2 statistic for which statistical significance was determined with the χ^2 distribution. For genome-wide data, the clustering algorithm would clearly be the analytical tool of choice because of the minimal computational resources required and the minimal time needed to carry out, in addition to its equivalent power to the slower permutation based threshold score.

We posited that the 2×2 tables would be more powerful than the $R \times 2$ tables using the clustering algorithm, since the 2×2 tables would decrease the dimensionality of the $R \times 2$ tables. However, the lesser degrees of freedom did not overcome the loss of information resulting from grouping all haplotypes with excess sharing into one cluster. Therefore, we would recommend the use of the $R \times 2$ tables in the clustering algorithm and not the 2×2 tables.

There were several limitations of this investigation. First, we assumed that the data was phased for which we initially phased the genotype data using fastPHASE. If phasing must be computationally inferred, errors in the estimation of haplotypes clearly could cause the results to be misleading if associations were observed. Furthermore, if haplotype sharing techniques would be applied on a genome-wide scale, genotype data that must be phased would be an immense computational endeavor. Future work to incorporate phasing of genotype data is undoubtedly needed.

Second, our results were limited to the unrelated case-control study design, though

analogous examinations on the power for other study designs could be conducted. In fact, much of the initial C code for the simulation engine and haplotype analysis was designed for the parent-parent-affected offspring trio design.

Third, we have presented the usefulness of haplotype sharing under the alternative hypothesis of a rare, highly penetrant, and recessive disease in which the disease harboring haplotypes emerged from founder populations. This is a very specific disease scenario amongst a spectrum of disease possibilities such as any combination of 1) the disease prevalence, e.g. from rare to common; 2) varying degrees of penetrance; 3) the genetic mode-of-inheritance risk models, e.g. multiplicative, dominant, over-dominant; 4) the presence or absence of a founder effect. In the initial stages of this investigation, we simulated a multiplicative genetic mode-of-inheritance risk model, no founder effect, and both common and rare causal SNPs sampled from the 6,170 phased iControlDB Caucasian haplotypes. The $R \times 2$ and 2×2 clustering algorithm nor the reference marker, window based, nor threshold scores exceeded the power of the allelic or haplotype χ^2 tests. Therefore, the haplotype sharing techniques that we have shown to have adequate power in this report would perform well under this particular disease setting (i.e. rare, highly penetrant, and recessive).

Lastly, we did not research the option of allowing for up to one mismatch on either side of the reference marker, as Lange and Boehnke (2004) did. Thus, future investigations could address the issues of genotyping error, recombination events, and mutations for haplotypes that are similar yet not completely identical though they may share a common ancestor.

In this report, we conducted a comprehensive investigation of haplotype sharing methods that have been proposed by other authors. It is our understanding of the current literature that to date there does not exist such a study. In addition, we elaborated on methods initially introduced by Lange and Boehnke (2004), such as the

use of thresholds with reference marker scores and the $\log_{10}$(CHSS) reference marker score, and we proposed a quick, efficient, and powerful algorithm that clusters like haplotypes. We found that the $\log_{10}$ version of the CHSS outperformed the other reference marker scores, dichotomizing the haplotype sharing scores with a threshold based on percentiles increased power, using fixed windows was detrimental to power, removing rare SNPs and SNPs in high LD with each other was not recommendable, and our novel clustering algorithm had competitive power and was significantly faster than permutation testing, which is desirable for genome-wide scans.

CHAPTER 4

GENE AND PATHWAY-BASED P-VALUES

4.1 Introduction

A commonly used approach in candidate gene association studies is to individually test each SNP within the candidate gene and focus attention on the most significant SNP or SNPs while ignoring those SNPs that did not reach statistical significance. However, this "most-significant SNP" approach can pose several problems. First, depending on the number of interrogated SNPs and hence the burden of multiple hypothesis testing, genetic variants that confer small disease risks could be missed. The "most-significant SNP" approach is not conducive to the phenomenon that multiple loci within a gene or multiple genes within a pathway often jointly function together in the etiology of common diseases (Peng et al., 2009). Second, due to locus heterogeneity (i.e. alleles at different loci that cause diseases in different populations), replicating a significant finding at a single marker can be a difficult task (Sladek et al., 2007). Replicating a significant association at the gene level may be easier than at the SNP level since a gene, particularly a pathway, comprises an interplay of components that act together

to perform specific biological tasks (Peng et al., 2009). Third, often the location of the specific causal variant is unknown and therefore we rely on indirect association with a SNP that is in high linkage disequilibrium (LD) with the actual susceptibility locus. Thus, a given single SNP may poorly predict the unobserved causal variant. Whereas a joint analysis of SNPs within a gene could be more powerful since information is combined across a collective number of SNPs.

Several methods have been suggested in order to jointly analyze SNPs within a gene or region, such as Fisher's method (Fisher, 1932), Hotelling's T^2 test (i.e. the standard multivariate test) (Xiong et al., 2002; Chapman et al., 2003; Fan and Knapp, 2003), a weighted Fourier transform (Wang and Elston, 2007), and conventional haplotype analysis (Schaid et al., 2002). Furthermore, in attempts to combine the information amongst the set of single SNP tests conducted, often the maximum value (i.e. minimum p-value) serves as the test statistic, for which the null distribution is unknown and its corresponding p-value is empirically evaluated through permutations of the disease status across all individuals (Chapman and Whittaker, 2008) or a conservative Bonferroni correction is applied. In gene expression microarray experiments, it is generally acceptable practice to represent the expression value of a gene by the maximum or median value of all its transcripts and/or probe sets, since a typical gene has only a few transcripts, and their expression levels are generally correlated (Wang et al., 2007). On the other hand, in a gene association study, a few or several hundred common SNPs on a chip may represent a typical gene, yet only one or a few of them contribute to disease susceptibility or are in LD with causal variants (Wang et al., 2007). Therefore, it is not immediately clear if taking the maximum test statistic proves to be a fair representation of the gene's impact on disease risk. This may be particularly true as focus shifts to rare variants.

Fisher's method for combining p-values is the following statistic,

$$T_{\text{Fisher}} = -2 \cdot \sum_{j=1}^{n} \log(p_j) \tag{4.1}$$

where p_j is the p-value corresponding to the single locus test at locus j and n is the number of loci. Fisher showed that under the global null hypothesis, T_{Fisher} follows a χ^2 distribution with $2n$ degrees of freedom granted that the tests are mutually independent. However, the single locus tests at nearby loci are likely to be correlated and therefore the limiting distribution of T_{Fisher} is unknown and statistical significance is estimated via permutation tests.

Assuming a single underlying causal locus and a genotype-based model, Chapman et al. (2003) showed that the appropriate multivariate score test statistic is

$$T_{\text{Hotelling}} = U^T V^{-1} U \tag{4.2}$$

where $U = \sum_{i=1}^{N} (Y_i - \bar{Y}) X_i = X^T (Y - \bar{Y})$, N is the number of genotyped individuals, $Y = (Y_1, \ldots, Y_N)^T$ is an $N \times 1$ vector of phenotypic values, $X = (X_1, \ldots, X_N)^T$ is a $N \times n$ matrix of genotype data for which each of N subjects contributes n genotypes across the genotyped loci (i.e. a $n \times 1$ vector of genotypes, X_i), and V is the estimated null variance-covariance matrix of U. Under the null hypothesis of no association between the putative causal locus and the phenotype, $T_{\text{Hotelling}}$ has an asymptotic χ^2 distribution with n degrees of freedom.

Multiple linear/logistic regression is also an alternative to testing multiple variant sites simultaneously. Such an approach could have substantial advantages over single SNP analyses (Balding, 2006). For example, covariates such as gender, age, environmental exposures, or SNP interactions can be included in the model. On the other hand, multiple regression-like analysis methods are appropriate only if the predictor

variables (e.g. the multiple marker loci) are not in strong LD, since an underlying assumption is that the predictor variables are independent. For example, if there are SNPs that are in LD with multiple functionally relevant variants, then standard regression analyses that do not control for the multicollinearity (i.e. the LD) would result in misleading results (Draper and Smith, 1981). Furthermore, in the presence of moderate to strong multicollinearity among the predictor variables, fitting the regression model could be computationally problematic when singular matrices must be inverted (Draper and Smith, 1981). Additionally, when the number of predictor variables far exceeds the number of subjects, such as in the case of genetic association studies for which thousands of SNPs may be genotyped, the least squares solution to estimation either can not be obtained or is highly problematic.

To circumvent the above-mentioned issues in multiple regression analysis of genetic association studies, Malo et al. (2008) propose the use of ridge regression. Ridge regression can deal with a large number of predictor variables compared to the number of subjects as well as predictors that are highly correlated. Ridge regression allows the inclusion of all SNPs in the model, rather than selecting a "representative" subset of SNPs as potential phenotype predictors. Since the 1970s, ridge regression has been available as a statistical tool to deal with multicollinearity, and small sample size and/or a large number of explanatory variables (Gruber, 1998; Hastie et al., 2001). Ridge regression places constraints on the size of the parameter estimates in attempts to control the large variances. In other words, these constraints effectively "shrink" the contribution of the redundant variables (e.g. the SNPs that are in strong LD with each other) toward zero. The ridge estimates of the usual linear regression model, $Y = X\beta + \varepsilon$ is

$$\hat{\beta}^{\text{Ridge}} = (X^T X + kI)^{-1} X^T Y \tag{4.3}$$

where X is an $n \times p$ matrix of genotypes (p is the number of SNPs or markers genotyped

on a set of n subjects), Y is an n-dimensional vector containing phenotype values for each individual, and the ridge parameter $k > 0$ represents the degree of shrinkage (Malo et al., 2008). The term kI aids in reducing multicollinearity and preventing the matrix X^TX from being singular even if X is not full rank. None of the regression coefficients are allowed to become very large, therefore it guards against over fitting and high variances usually associated with correlated coefficients. In contrast, the standard parameter estimates obtained by minimizing the residual sum of squares, $RSS = (Y - X\beta)^T(Y - X\beta)$, is

$$\hat{\beta}^{\text{MLR}} = \arg\min(RSS) = (X^TX)^{-1}X^TY \tag{4.4}$$

Wu et al. (2009) suggest the use of lasso penalized logistic regression for genome-wide association analysis. The lasso penalty is another effective device for continuous model selection, particularly when the number of predictors p far exceeds the number of observations n (Tibshirani, 1996). Let y_i be the response for case i, x_{ij} the j-th predictor for case i, β_j the regression coefficient corresponding to the j-th predictor and μ the intercept. Also let $\theta = (\mu, \beta_1, \ldots, \beta_p)^T$ and $x_i = (x_{i1}, \ldots, x_{ip})^T$. The objective function in ordinary linear regression is $f(\theta) = \sum_{i=1}^{n}(y_i - \mu - x_i^T\beta)^2$, whereas in lasso penalized regression, the following modified objective function is minimized

$$g(\theta) = f(\theta) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{4.5}$$

where the tuning constant λ controls the strength of the penalty, which shrinks each β_j toward the origin and enforces sparse solutions (Wu et al., 2009).

As an alternative to penalized regression for model selection in order to jointly assess statistical significance for a set of SNPs within a gene of interest, the authors of PLINK (Purcell, 2007; Purcell et al., 2007) implement an algorithm based on pruning SNPs

in LD with each other as measured by r^2. Under the section "Association / Set-based tests", the authors describe the algorithm in the following manner:

1. For each set, for each SNP determine which other SNPs are in LD, above a certain threshold R.

2. Perform standard single SNP analysis (which might be basic case/control association, family-based TDT or quantitative trait analysis).

3. For each set, select up to N "independent" SNPs (as defined in step 1) with p-values below P. The best SNP is selected first; subsequent SNPs are selected in order of decreasing statistical significance, after removing SNPs in LD with previously selected SNPs.

4. From these subsets of SNPs, the statistic for each set is calculated as the mean of these single SNP statistics.

5. Permute the dataset a large number of times, keeping LD between SNPs constant (i.e. permute phenotype labels).

6. For each permuted dataset, repeat steps 2 to 4 above.

7. Empirical p-value for set (EMP1) is the number of times the permuted set-statistic exceeds the original one for that set.

There are potential problems with this strategy of pruning SNPs based on the r^2 measure of LD. If many SNPs are correlated, the chosen SNP may not actually be the functional SNP. Furthermore, it is possible that more than one SNP is functional amongst those that are in moderate LD, such that choosing one to represent a cluster of correlated SNPs would not reflect the fact that more than one position in the sequence is phenotypically relevant (Malo et al., 2008). Perhaps of greater concern is the forced

evaluation of a single test statistic defined by the mean of all individual SNP test statistics for individual "independent" (independence defined based on a user-defined *ad hoc* threshold) SNPs that reach a user-defined threshold of statistical significance. In this manner, it is plausible that a highly significant SNP will be combined with several marginally significant results and the resulting test statistic would likely reflect a far less significant finding than if the most significant SNP was the only SNP considered. Furthermore, the iterative inclusion of individual SNP test statistics to the running sum (from all other previously included SNPs) test statistic representing the combined effect of the most significant SNPs ignores the impact of modest linkage disequilibrium between the SNPs on the individual SNP test statistics.

Li and Leal (2008) propose two strategies for combining information across multiple marker loci in the presence of rare variants. Li and Leal discuss their methods in light of the next generation sequencing efforts that would inevitably lead to the identification of rare variants, which will comprise both nonfunctional and functional rare variants in disease etiology. They argue that an effective first approach is to identify the genes that are involved in disease onset, although understanding the effects of specific rare variants is ultimately important. The details of their proposed methods are the following. Assume there are N cases and N controls that are genotyped across M SNPs. For Li and Leal's "Collapsing Method", they set an indicator variable X for the j-th case individual to 1 if rare variants are present within the M SNPs and 0 otherwise. They define Y_j similarly for the control individuals. The authors claim that due to the rarity of variants, the probability of carrying more than one variant for an individual is low. Thus, collapsing the genotypes across all variants could enrich the association signals, granted that nonfunctional variants are not intermingled with functional ones. Li and Leal test whether the proportion of individuals with rare variants in the cases (ϕ_A) and controls ($\phi_{\bar{A}}$) differ, i.e. $H_0 : \phi_A = \phi_{\bar{A}}$, by way of Pearson's χ^2 statistic.

Moreover, Li and Leal (2008) propose the "CMC Method" that combines collapsing and multivariate tests. Based on some predefined criteria such as marker allele frequencies, the M markers are divided into k subgroups $(g_1, \ldots, g_k)$ for which each group g_j contains n_j $(j = 1, \ldots, k)$ SNP members. For each group g_j, the Collapsing Method as described above is carried out such that each subject has k indicator variables X that specify the presence or absence of at least one rare variant within the group g_j. Note that no collapsing is performed for groups with a single member. Lastly, Hotelling's T^2 test is utilized on the resulting data structure to jointly assess the statistical significance of the gene.

Li et al. (2009) introduce yet another gene-based association test (ATOM: a multi-marker Association Test by combining Optimally weighted Markers) that incorporates marker weights that are proportional to the amount of information it captures about the unknown trait locus. In particular, Li et al. define a score for an individual i

$$S_i = \frac{1}{m} \sum_{j=1}^{m} w_j g_{ij} \tag{4.6}$$

where m is the number of markers under study, w_j weights the genotype at marker j, and g_{ij} is the genotype at marker j for subject i for which $g_{ij} \in \{0, 1, 2\}$ counts the number of alleles 1_j (markers are diallelic with alleles 1_j and 0_j). The genotype weights are defined as

$$w_j = \frac{\Delta_j}{p_j q_j} \tag{4.7}$$

where p_j and q_j are the allele frequencies at marker j and Δ_j is the LD coefficient between the quantitative trait locus (QTL) and marker j (typically referred to as D and computed as $\Delta_j = p_T 1_j - p_T p_j$ where p_T is the allele frequency of T at the diallelic QTL and $p_T 1_j$ is the joint probability of alleles T and 1_j at the QTL and marker j, respectively. Δ_j measures the difference between the observed joint frequency of T

and 1_j and the expected co-occurrence of T and 1_j assuming independent loci). The authors motivate w_j by proving that the beta coefficients in the linear regression models that regress the QTL genotypes on the phenotype and the marker genotypes on the phenotype follow the relationship

$$\beta = \beta_T \frac{\Delta}{p_A p_a} \tag{4.8}$$

where β and β_T are the regression coefficients of the genotypes for the marker and QTL, respectively, p_A and p_a are the allele frequencies at a given marker for alleles A and a, respectively, and Δ is the LD coefficient between the QTL and marker. Thus, the slopes at the marker and QTL differ by a factor $\Delta/(p_A p_a)$, the weighting function. In general, the stronger the LD is between the trait and marker loci, the greater the magnitude will be for the weight w_j. Therefore, the score S_i defined in Equation 4.6 effectively allocates weights to markers according to their levels of LD with the trait locus (Li et al., 2009). The association information contained in all m markers is captured in one score for an individual i and the dimension is reduced from m to 1.

Since the location of the QTL is unknown, the weights must be estimated. Li et al. (2009) propose gathering information on the LD structure of the candidate gene by using a reference dataset such as the genotypes contained in the International HapMap Project (www.hapmap.org), other publicly available dense SNP datasets, or resequencing data from a subset of the study sample. In particular, employing the reference sample containing M markers as well as the study sample consisting of m markers (for which $M > m$ in most scenarios) at each marker k in the reference dataset ($1 \leq k \leq M$) would yield a score $S_{i,k}$ for every individual i

$$S_{i,k} = \frac{1}{m} \sum_{j=1}^{m} w_j^k g_{ij} \tag{4.9}$$

where $w_j^k = \Delta_j^k / p_j q_j$ is the LD coefficient between markers k and j and p_j and q_j are the allele frequencies at marker j. Upon computing these score estimates for all of the subjects in the study sample, each subject will have M scores. Li et al. (2009) then conduct principal components analysis (PCA) in order to reduce the dimensionality of the dataset while retaining as much as possible the variation contained therein. The PCA transforms the original set of M correlated scores $\{S_{1,k}, \ldots, S_{n,k}\}_{k=1}^{M}$ across n individuals into a set of m uncorrelated principal components. Once the principal components are computed, a series of linear (for a quantitative phenotype) or logistic (for a binary trait) regression models are constructed in which the predictive set of principal components for a given model is selected based on a designated proportion of the variance explained by the principal components. For each model, a joint test involving all the regression coefficients is conducted. The authors choose the maximum statistic, T_{ATOM}, as the test statistic and estimate its statistical significance via permutation tests.

Li et al. (2009) suggest that their method is different from traditional PCA-based approaches that operate directly on the marker genotypes observed in the study sample. In contrast, their method operates on the set of scores $\{S_{1,k}, \ldots, S_{n,k}\}_{k=1}^{M}$ for all M markers in the reference dataset. With traditional PCA approaches, the regression coefficients are determined solely by the correlation structure among the genotyped markers. Whereas with Li et al.'s strategy, additional LD information for that region is integrated into their defined weights w_j^k.

We propose a method based on forward variable selection in regression that provides a joint test of the statistical significance of a gene. We compared our method with existing and conventional methods such as computing the minimum p-value while assessing statistical significance via permutation testing and PLINK's Set-Based Association Test that calculates the average test statistic for a set of single SNPs and

the overall p-value is also determined via permutation testing. We evaluated all methods under various alternative hypotheses, simulating candidate gene studies as well as studies involving a candidate list of genes. Each method excelled in certain circumstances, for example, when simulating candidate gene studies, the minimum p-value based approaches appeared to be more powerful under alternative models for which a single QTL was responsible for the genetic variation. Both PLINK and our method were most powerful in detecting more than one QTL, and our method performed the best when pairwise SNP x SNP interactions were modeled. Lastly, for the simulations comprising multiple candidate genes, our approach and Fisher's Method had consistent and greater power than the other techniques.

We note that although an aspect of our simulation study consisted of analyzing a set of various candidate genes, one could plausibly apply the foregoing analytical methods to a putative biological pathway comprising several genes.

4.2 Methods

4.2.1 Minimum P-Value Across All SNPs

We employed the approach of computing the minimum p-value across all of the single SNP tests, as discussed in the Introduction (Section 4.1). We estimated the statistical significance of this minimum p-value by permuting a large number of times the column vector of quantitative phenotypes in a given simulated data set (while preserving the LD structure of the SNPs across the sample of simulated subjects), and for each permutation computing the minimum p-value from the single SNP tests. The permutation adjusted minimum p-value was then taken to be the proportion of times the permuted minimum p-values were less than or equal (i.e. as extreme or more extreme) than the observed minimum p-value.

For a data set containing SNPs within a specific gene region, this method offered a way to test the significance of the gene. Furthermore, for data sets comprising multiple genes within a biological pathway, the resulting p-value was an estimate of the statistical significance of the pathway.

4.2.2 Minimum P-Value by Gene, Bonferroni Adjusted

Across N_{Genes} genes within a given data set, we calculated the minimum p-value amongst the single SNP tests within a particular gene, resulting in N_{Genes} gene-specific minimum p-values. We then estimated the statistical significance of these minimum p-values via permutation testing as described above in Section 4.2.1. To adjust for multiple hypothesis testing, we then applied the Bonferroni correction to the minimum of the permutation adjusted minimum gene-specific p-values. Namely, we multiplied the overall minimum p-value by N_{Genes} and simply took the p-value to be one if this product exceeded one.

We note that for data sets comprising a single gene, this method was equivalent to taking the minimum p-value across all single SNP tests discussed in Section 4.2.1.

4.2.3 Minimum P-Value by Gene, Fisher's Method

As an alternative to the Bonferroni adjustment detailed above in Section 4.2.2, we carried out Fisher's method for combining p-values, described in the Introduction (Section 4.1) and contained in Equation 4.1. We note that for single gene data sets, this was not equivalent to computing the minimum p-value amongst all single SNP tests (Section 4.2.1) since T_{Fisher} was ultimately based on a χ^2 distribution with $2N_{\text{Genes}}$ degrees of freedom.

4.2.4 PLINK's Set-Based Association Test

We performed PLINK's set-based association test, explicitly described in the Introduction (Section 4.1). For data sets consisting of a single gene or entire biological pathway, we defined the set to be the gene or pathway. We used the most current version (1.06) of PLINK available at the time of implementation (Purcell, 2007; Purcell et al., 2007).

4.2.5 Joint Test Based on Forward SNP Selection

We propose a method based on forward variable selection in regression that provides a joint test of the statistical significance of a gene or pathway. We assume that we have n subjects that are genotyped on p SNPs within a candidate gene or pathway and that have been measured for a quantitative phenotype. Specifically, the algorithm to build the multi-SNP linear regression model is the following, for $j = 1, \ldots, p$ SNPs:

1. Begin forward selection (i.e. when $j = 1$) by carrying out single SNP linear regression models across all p SNPs. After the first iteration (i.e. for $j > 1$), adjust for the selected SNPs (i.e. $SNP_1, \ldots, SNP_{j-1}$) in the model.

2. While adjusting for the selected set of SNPs ($SNP_1, \ldots, SNP_{j-1}$), construct $p-(j-1)$ linear regression models across the remaining $p-(j-1)$ SNPs that have not yet been selected to represent the candidate gene or pathway. Do not consider any SNPs that are in "high" LD with any of the SNPs already entered in the SNP covariate set, based on a user-defined r^2 threshold (i.e. prune any SNPs with pairwise r^2 values *above* a given r^2 threshold). Note, for $j = 1$ we simply select the SNP with the smallest p-value.

3. For each of the $p-(j-1)$ models, conduct a joint test of all the SNPs, i.e. $H_0 : \beta_1 = \cdots = \beta_j = 0$ and find the most significant p-value. If this minimum joint p-value is *smaller* than the $j-1$-th joint p-value, then add this SNP to the SNP covariate set,

provided that the p-value corresponding to the test on the individual parameter is *less than* a user-defined p-value threshold. If the individual p-value does not meet this p-value threshold, then consider the next "most significant" joint (that also improves upon the prior joint p-value) and corresponding individual p-values, and so forth. As a final filter, the SNP covariate set may not exceed a user-defined number of members. The j-th SNP corresponding to this joint test is selected as a predictor from the candidate gene or pathway in explaining the phenotypic variation. Record the p-value for this joint test under the j-th iteration.

4. Repeat steps 1 through 3 for each iteration of j until no more SNPs can be added in the multi-SNP linear regression model, based on the predefined stopping criterion defined in the prior step. When the forward selection procedure ceases, there will be p^* SNP predictors in the model that contains the largest number of variables.

5. The minimum p-value amongst the p^* joint p-values will be the last joint p-value recorded, as defined by the nature of the algorithm. This set of $p^*_{\min}$ SNPs is chosen to act as a proxy for the candidate gene or pathway. Estimate the p-value of this test statistic via permutation testing. This joint p-value represents the statistical significance of the candidate gene or pathway.

Alternative Stopping Criterion

We allowed for a more relaxed criterion in building the SNP covariate set. If the current joint p-value being evaluated did not improve upon the prior joint p-value, then we admitted this SNP in the set (granted that its individual p-value met the threshold and the maximum number of SNP members in the set was not yet satisfied) and continued building the set under the usual guidelines as specified above. We ceased

to expand the SNP set when we encountered a joint p-value that was not smaller than the overall minimum joint p-value.

Inclusion of Pairwise SNP x SNP Interactions

We designed the option to include pairwise SNP x SNP interactions. As we constructed the SNP covariate set, we sequentially added all possible pairwise SNP x SNP interactions in the linear model containing the current state of the SNP set (as well as all other previously entered interactions). We decided to keep the interaction term if the joint p-value that assessed all terms in the model was more significant.

A Note on the Thresholds: r^2, Individual P-Value, and Max Number of SNP Members

We analogously implemented the r^2, individual p-value, and maximum number of SNP members thresholds described in the Introduction in reference to PLINK's set-based association test (Section 4.1) so as to allow a fair and direct comparison of PLINK's approach and our competing method. The essential difference between the two techniques was that PLINK assessed overall statistical significance by averaging the single SNP test statistics contained in the set, whereas the p-value in our proposed method was based on the joint test of the parameters in a general linear model.

Setting the maximum number of members in a set to one and not imposing a p-value filter (i.e. setting the p-value threshold to one) resulted in a test based on the "best" single SNP for PLINK's set-based test and our forward selection procedure. On the other hand, by not constraining the number of SNPs in the set and by turning the p-value and r^2 filters off (i.e. p-value threshold = 1 and r^2 threshold = 1), PLINK's test included *all* test statistics across all of the SNPs in the data set. For our method, it was not feasible to construct a regression model with a considerable number of parameters.

4.2.6 Simulations: FTO as a Candidate Gene

To evaluate the power of the proposed methods, we simulated genotype-phenotype data for a candidate gene in the following manner. We designated FTO (official full name: fat mass and obesity associated; residing on chromosome 16; 410.5 kilobases in length; spanning base pair positions 52,295,376 to 52,705,882) as the candidate gene and subsetted the corresponding genotypes in 3,172 Caucasian subjects from Illumina's iControlDB who were genotyped on Illumina's HumanHap550 platform. These subjects served as the sampling pool from which we generated simulated data sets. There were a total of 97 representative SNPs in FTO on this genotype panel. For the purposes of the simulation, we required all of the subjects to have been successfully genotyped at all 97 SNP loci. Thus, we removed 510 subjects who were missing at least one genotype, leaving a total of 2,662 subjects in the genotype sampling pool.

Based on a review of the literature, we chose SNP rs8050136 within FTO to act as a "causal" locus. This SNP has been previously reported in investigations on obesity (Scott et al., 2007; Scuteri et al., 2007; Grant et al., 2008; Thorleifsson et al., 2009; Pecioska et al., 2010) and is a perfect proxy to rs9939609, for which we discerned using the web-based tool from the Broad Institute, SNAP (SNP Annotation and Proxy Search; Johnson et al., 2008; http://www.broadinstitute.org/mpg/snap/), based on the CEU data in HapMap. The SNP rs9939609 is located within FTO and has been widely published (Frayling et al., 2007; Scuteri et al., 2007; Wellcome Trust Case Control Consortium, 2007).

To simulate the data for the power analyses, we randomly picked 3,000 subjects (with replacement) from the iControlDB sampling pool described above. We assumed that the quantitative phenotype was normally distributed and for each of the 3,000 simulated Caucasian subjects we generated the quantitative phenotype by drawing a normal random variate based on the genotypes observed at the rs8050136 "suscepti-

bility" locus for that subject. Specifically, we assumed that rs8050136 explained 0.5% of the phenotypic variation (i.e. we set the coefficient of determination of the simple linear regression model, R^2, to 0.005) and then solved for the slope parameter using the relationship

$$\beta_1 = \sqrt{\frac{\sigma_Y^2}{\sigma_X^2} \cdot R^2} \tag{4.10}$$

where σ_Y^2 and σ_X^2 were the variance of the quantitative phenotype and genotypes (at rs8050136, in our example), respectively. We defined the variance of the phenotype to be $\sigma_Y^2 = 1$ and its mean to be $E(Y) = 0$, i.e. the phenotype followed a standard normal distribution. Assuming an additive genetic model coding for the bi-allelic SNP, the expectation and variance of the genotypes were

$$E(X) = \mu_X = 0 \cdot f_{dd} + 1 \cdot f_{Dd} + 2 \cdot f_{DD} \tag{4.11}$$

$$V(X) = \sigma_X^2 = (0 - \mu_X)^2 \cdot f_{dd} + (1 - \mu_X)^2 \cdot f_{Dd} + (2 - \mu_X)^2 \cdot f_{DD} \tag{4.12}$$

where $X = 0, 1, 2$ under the additive coding for genotypes dd, Dd, and DD, respectively, and f_{dd}, f_{Dd}, and f_{DD} were the probabilities of observing said genotypes. We assumed Hardy-Weinberg Equilibrium such that $f_{dd} = f_d^2$, $f_{Dd} = 2f_D f_d$, and $f_{DD} = f_D^2$, where f_d and f_D were the frequencies of the non-risk and risk alleles, respectively. Finally, since the mean of the quantitative phenotype was zero, then the intercept was

$$\begin{aligned} E(Y \mid \mu_X) = 0 &= \beta_0 + \beta_1 \cdot \mu_X \\ \beta_0 &= -\beta_1 \cdot \mu_X \end{aligned} \tag{4.13}$$

where β_1 and μ_X were defined in Equations 4.10 and 4.11, respectively.

Therefore, utilizing the relationships specified above in Equations 4.10 through 4.13, for rs8050136, $f_D = 0.3989$ in the iControlDB genotype sampling pool, $\sigma_Y^2 = 1$, and $R^2 = 0.005$, the intercept and slope were computed to be -0.08146 and 0.1021, respec-

tively.

Using this information, upon constructing the data set for the simulations, the quantitative phenotype for each of the 3,000 randomly sampled vectors of genotypes was generated by randomly sampling from the normal distribution with mean $\beta_0 + \beta_1 X$ and standard deviation of one, where $X(= 0, 1, 2)$ was the observed genotype at rs8050136.

Single SNP Simulation Models for SNPs in Low to Modest LD with rs8050136

In addition to the data sets described above for which rs8050136 was the causal locus, we also generated an additional two sets of data sets using alternative causal loci. In the first and second sets, we chose SNPs in low and modest LD with rs8050136 (MAF = 0.3989), rs16953002 had a pairwise $D' = 0.056$ (MAF = 0.1741) and rs10521307 had a pairwise $D' = 0.311$ (MAF = 0.2977). We subsequently generated the two sets of data sets with rs16953002 and rs10521307 as the causal loci, in the same manner described above for rs8050136. The corresponding intercept and slope for the rs16953002 models were -0.04591 and 0.1319, and for rs10521307 were -0.06511 and 0.1094.

Two SNP Simulation Models for SNPs in Low to Modest LD with rs8050136

We simulated two other sets of data sets for which two SNP loci were responsible for the genetic variation observed in the quantitative trait. The first SNP was the original rs8050136 SNP and the second SNP was either rs16953002 (low LD with rs8050136) or rs10521307 (modest LD with rs8050136). The effect sizes (i.e. slope parameters) were defined as above for the *single* SNP data sets, such that the two causal SNPs explained a cumulative proportion of 1% of the total phenotypic variation. In contrast to that described above for the *single* SNP data sets, to generate the quantitative trait we randomly sampled from a normal distribution with mean $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ and standard deviation of one, where X_1 was the sampled genotype (0, 1, or 2) at

rs8050136 and X_2 was the sampled genotype at either rs16953002 or rs10521307.

Simulation Models Including Pairwise SNP x SNP Interactions with rs8050136

Building upon the *two* SNP models described above, we added pairwise SNP x SNP interactions with rs8050136. To accomplish this, we simply incorporated an interaction parameter when randomly sampling from the normal distribution with mean $\beta_0+\beta_1 X_1+ \beta_2 X_2 + \beta_3 X_1 X_2$ and standard deviation of one. $X_1 X_2$ was the product of the observed sampled genotypes, e.g. $X_1 X_2$ could have the values of 0, 1, 2, or 4.

The specification of the interaction parameter was more involved than for the simple linear regression models. We assumed that the total phenotypic variation could be decomposed into both genetic and environmental components,

$$\sigma_T^2 = \sigma_G^2 + \sigma_E^2 \tag{4.14}$$

For example, by assuming that the quantitative trait arose from a normal distribution with standard deviation of one, then $\sigma_T^2 = 1$. In addition, for these interaction models, we assumed that the total *genetic* variation was fixed at 1%, i.e. $\sigma_G^2 = 0.01$. Since we further assumed the mean of the phenotype to be zero ($E(Y) = 0$), then from the definition of variance,

$$\begin{aligned} \sigma_T^2 &= E(Y^2) - [E(Y)]^2 \\ &= E(Y^2) \end{aligned} \tag{4.15}$$

Therefore, the genetic variation for a two SNP model including interactions (first SNP:

minor/major alleles a/A; second SNP: minor/major alleles b/B) was

$$\begin{aligned}
\sigma^2_G = \quad & f_{aa}f_{bb}(\beta_0 + 2\beta_1 + 2\beta_2 + 4\beta_3)^2 \\
& + f_{aa}f_{Bb}(\beta_0 + 2\beta_1 + \beta_2 + 2\beta_3)^2 \\
& + f_{aa}f_{BB}(\beta_0 + 2\beta_1)^2 \\
& + f_{Aa}f_{bb}(\beta_0 + \beta_1 + 2\beta_2 + 2\beta_3)^2 \\
& + f_{Aa}f_{Bb}(\beta_0 + \beta_1 + \beta_2 + \beta_3)^2 \\
& + f_{Aa}f_{BB}(\beta_0 + \beta_1)^2 \\
& + f_{AA}f_{bb}(\beta_0 + 2\beta_2)^2 \\
& + f_{AA}f_{Bb}(\beta_0 + \beta_2)^2 \\
& + f_{AA}f_{BB}(\beta_0)^2
\end{aligned} \tag{4.16}$$

where f_{aa}, f_{Aa}, and f_{AA} and f_{bb}, f_{Bb}, and f_{BB} were the probabilities of observing the genotypes for SNPs one and two. We note that we assumed the two SNPs to be completely independent. Similarly, for $E(Y) = 0$

$$\begin{aligned}
E(Y) = 0 = \quad & f_{aa}f_{bb}(\beta_0 + 2\beta_1 + 2\beta_2 + 4\beta_3) \\
& + f_{aa}f_{Bb}(\beta_0 + 2\beta_1 + \beta_2 + 2\beta_3) \\
& + f_{aa}f_{BB}(\beta_0 + 2\beta_1) \\
& + f_{Aa}f_{bb}(\beta_0 + \beta_1 + 2\beta_2 + 2\beta_3) \\
& + f_{Aa}f_{Bb}(\beta_0 + \beta_1 + \beta_2 + \beta_3) \\
& + f_{Aa}f_{BB}(\beta_0 + \beta_1) \\
& + f_{AA}f_{bb}(\beta_0 + 2\beta_2) \\
& + f_{AA}f_{Bb}(\beta_0 + \beta_2) \\
& + f_{AA}f_{BB}(\beta_0)
\end{aligned} \tag{4.17}$$

We had two equations (Equations 4.16 and 4.17) and two unknowns (β_0 and β_3) for which β_1, β_2, and the genotype probabilities were completely specified. We solved for β_0 in Equation 4.17 and substituted β_0 into Equation 4.16. This resulted in a quadratic equation which we solved algebraically using the quadratic formula. Thus, given the two SNPs' allele frequencies, effect sizes, and genetic variation, we were able to solve for the interaction parameter and intercept of the multiple linear regression model.

Namely, for the simulated data sets modeling rs8050136, rs16953002 (low LD with rs8050136), and their pairwise interaction, the interaction parameters were 0.05659 and -0.2331 with corresponding intercepts -0.1058 and -0.02533. For the simulated data sets modeling rs8050136, rs10521307 (modest LD with rs8050136), and their interaction, the interaction parameters were 0.04185 and -0.1876 with corresponding intercepts -0.1235 and -0.01451.

Simulation Parameters and r^2, P-Value, and Max Number of SNPs in Set Filters

For the power analyses, we simulated 100 data sets for each simulation scenario in which each data set contained 3,000 subjects. Whereas for the type I error analysis, we simulated 500 data sets. For each data set, we permuted the phenotype 500 times in carrying out the permutation-based tests.

For PLINK's set-based association test and our forward selection procedure, we set the r^2 pruning thresholds to 1 (i.e. effectively no filter), 0.8, 0.5, and 0.1, the p-value thresholds to 1 (i.e. effectively no filter), 0.05, and 0.0025. The 0.0025 p-value threshold was chosen because based on prior power calculations, this cutoff was predicted to give us 80% power to detect a variant that explained 0.5% of the variation in 3,000 subjects. Lastly, the number of SNPs in the set was not allowed to exceed 5 or 10.

4.2.7 Simulations: Body Mass Index Related List of Candidate Genes

We assessed the power of the competing approaches in analyzing a list of candidate genes, in contrast to one specific candidate gene as described in Section 4.2.6. We continued under the hypothetical scenario of studying obesity as a quantitative trait such as body mass index (BMI). Willer et al. (2009) reported the association of SNPs within or near eight genes (NEGR1, TMEM18, GNDPA2, MTCH2, SH2B1, FTO, MC4R, and KCTD15). We note that the FTO gene (the gene we used to simulate the candidate gene studies) was also listed amongst these genes.

We prepared the pool of genotypes across the candidate genes in the following way. Within the iControlDB Caucasian sample (N = 3,172) and for each gene, we subsetted the SNPs located within 50 kilobase pairs upstream or downstream of the gene, including the SNPs in the gene itself, resulting in eight pools of gene-specific genotypes. Then, for each genotype pool we removed any subject that was missing at least one genotype at any of the loci. We were left with 1,663, 2,881, 3,135, 3,150, 3,036, 2,497, 2,993, and 2,677 vectors of genotypes for the NEGR1, TMEM18, GNDPA2, MTCH2, SH2B1, FTO, MC4R, and KCTD15 gene pools. The number of SNPs and other characteristics of these genes are reported in the Results (Section 4.3).

The flanking SNPs for each gene defined the boundaries for the minimum p-value gene-based methods described in Sections 4.2.2 and 4.2.3.

To generate the simulated data sets for the subsequent power analysis, we designated five "causal" SNPs and proceeded in the same fashion as that described for the two SNP candidate gene models (Section 4.2.6). The only difference being that instead of randomly sampling with replacement from one pool of genotypes, we selected from eight pools of gene-specific genotypes. The random selection of vectors of genotypes for one pool was independent of the random selection for another pool. In essence, we

assumed the genes were not in LD and so we did not attempt to preserve the observed LD structure across the genes.

We simulated data under two alternative models. For both models, we assigned two SNPs in FTO (rs8050136 and rs10521307, in moderate LD with the former) and one SNP near MC4R (rs10871777) as QTLs. We selected the two SNPs in FTO based on one of the two SNP causal models from the FTO candidate gene simulation study. In addition, we chose these two genes and SNPs because Willer et al. (2009) strongly confirmed their associations with BMI. We note that Willer et al. (2009) reported the variant rs17782313, though this SNP was not on the Illumina HumanHap550 panel. We therefore employed SNAP (Johnson et al., 2008) and located rs10871777, a perfect proxy ($r^2 = 1$ and $D' = 1$) for rs17782313 that was on the array and is 666 base pairs away. We allowed the primary FTO SNP (rs8050136) and the MC4R SNP to explain 0.35% of the phenotypic variation, whereas the other FTO SNP (rs10521307) explained slightly less ($R^2 = 0.25\%$). This corresponded to slope parameters of 0.08543, 0.07732, and 0.09836, respectively. Also, for both alternative models we simulated 3,000 subjects as we did prior for the FTO candidate gene models (Section 4.2.6), and an overall phenotypic mean and standard deviation of 0 and 1, respectively. We did not model any pairwise SNP interactions.

Under the first model, we chose two other SNPs (beyond the three described above) in MTCH2 (rs10838738) and SH2B1 (rs7498665), two genes for which there were few representative SNPs (nine and five) on the Illumina HumanHap550 genotyping platform. In contrast, under the second model, instead of the MTCH2 and SH2B1 causal SNPs, we selected a SNP from NEGR1 (rs2568958) and TMEM18 (rs4854344), two genes that contributed a fair number of SNPs to the analysis (144 and 34). Similar to the MC4R SNP, we identified proxies for the NEGR1 and TMEM18 SNPs as described in Willer et al. (2009), specifically, Willer et al. documented associations for rs2815752

(in NEGR1) and rs6548238 (in TMEM18), although these SNPs were not on the Illumina HumanHap550 array. We pursued the second model in order to assess the impact on power when the QTLs were embedded in a larger set of SNPs. We posited that the minimum p-value gene-based methods (for which the analyst must define the gene regions) would suffer in power compared to PLINK's or our step-wise method. For these secondary SNPs, we set R^2 to 0.25%, marginally less than the FTO and MC4R primary SNPs ($R^2 = 0.35\%$). The corresponding slope parameters of the MTCH2 and SH2B1 SNPs for the first model were 0.07460 and 0.07353, and under the second model the effect sizes of the NEGR1 and TMEM18 SNPs were 0.07501 and 0.09115, respectively.

Lastly, the vast majority of genes contained merely one or two SNPs on the HumanHap550 panel, thus we found it reasonable to include a 100 kilobase pair window about the genes of interest, so as offer a fair assessment of power for the gene-based minimum p-value methods with PLINK and our step-wise forward SNP selection procedure.

4.2.8 Computational Details

All code was written in R (R Development Core Team, 2006) and all power analyses were carried out on UNC's Topsail, a 4,160-processor Dell Linux cluster (2.3 GHz Intel EM64T processors and 12 GB of memory). Due to the computationally intensive permutation testing, we implemented parallel code in R using the library "snowfall", designating at least 8 CPUs for every simulated data set.

The LD displays were produced using Haploview version 4.1 (Barrett et al., 2005).

4.3 Results

Figure 4.1 contains the LD plot (D') of the FTO gene region in the iControlDB sample of Caucasians (N = 2,662 subjects with no missing genotypes at any SNP loci).

There were 97 representative SNPs on the Illumina HumanHap550 genotype platform. The primary causal SNP (rs8050136) was located at position number 12, amongst a relatively medium-sized LD block. The second disease bearing SNP was rs16953002, which was SNP number 87, positioned on the extreme downstream end of the FTO gene region. This SNP was in "low" LD with rs8050136 ($D' = 0.056$) and was seated in between relatively large (upstream) and small (downstream) LD blocks. The third quantitative trait locus selected was rs10521307 (position number 25) and was in "modest" LD with rs8050136 ($D' = 0.311$).

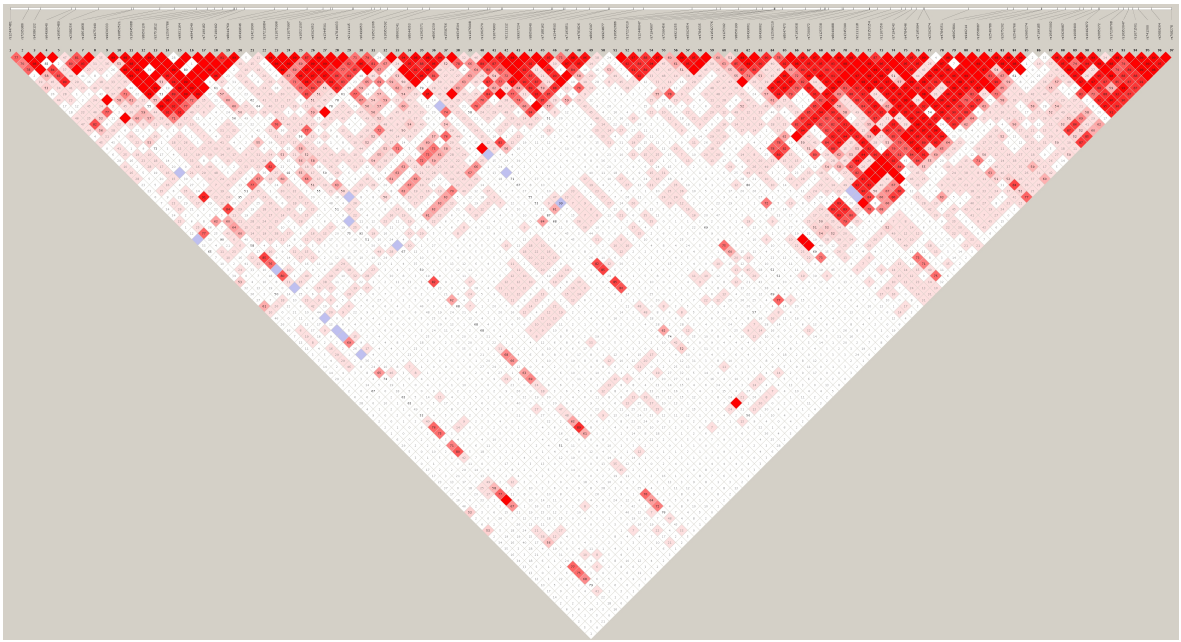

FIGURE 4.1: LD plot (D') of the FTO gene (residing on chromosome 16; 97 representative SNPs on the Illumina HumanHap550 genotype platform; 395.96 kilobases in length; spanning base pair positions 52,306,470 to 52,702,426) in the iControlDB sample of Caucasians (N = 2,662 subjects with no missing genotypes at any SNP loci). White represents $D' < 1$ and LOD < 2, shades of pink/red $D' < 1$ and LOD ≥ 2, blue $D' = 1$ and LOD < 2, and bright red $D' = 1$ and LOD ≥ 2.

Table 4.1 presents a brief characterization of the eight genes in the candidate gene list power analyses. The largest genes were NEGR1 and FTO and hence these genes correspondingly offered the most SNPs on the Illumina HumanHap550 array. On the

TABLE 4.1: Characterization of the eight genes selected for the candidate gene list analytical approach.

				Number of SNPs[a]				
			Size	Gene	50 kbp	"Risk" SNPs[c]		
	Gene	Chr	(kbp)	Only	Window[b]	Model 1	Model 2	MAF
1	NEGR1	1	879.7	135	144		rs2568958	0.3496
2	TMEM18	2	9.5	2	34		rs4854344	0.1836
3	GNPDA2	4	24.4	1	7			
4	MTCH2	11	25.2	2	9	rs10838738		0.3410
5	SH2B1	16	10.2	1	5	rs7498665		0.3628
6	FTO	16	410.5	97	110	rs8050136	rs8050136	0.3958
						rs10521307	rs10521307	0.2982
7	MC4R	18	1.4	1	15	rs10871777[d]	rs10871777[d]	0.2371
8	KCTD15	19	17.4	1	24			

Chr: Chromosome
kbp: kilobase pairs
MAF: Minor Allele Frequency
rs2568958 (in NEGR1), rs4854344 (in TMEM18), and rs10871777 (near MC4R) were strong proxies to rs2815752, rs6548238, and rs17782313, SNPs that Willer et al. (2009) previously reported to show significant evidence for association with BMI
[a] Based on the Illumina HumanHap550 genotyping platform
[b] We included SNPs residing within a 50 kilobase pair window flanking the gene
[c] We set $R^2 = 0.35\%$ for rs8050136 (in FTO) and rs10871777 (near MC4R) and $R^2 = 0.25\%$ for the remainder of the SNPs in the model
[d] rs10871777 is located *near* MC4R, not within

other hand, the smallest gene (MC4R) was not represented on the array, as well as the SNP rs17782313 that was reported to be associated with BMI (Willer et al., 2009). The variants rs2815752 (in NEGR1) and rs6548238 (in TMEM18) were also not on the chip, thus we chose strong proxies that were proximal to the reported causal SNPs. All of the SNPs included in the alternative models were fairly common (MAF above 18%).

Figure 4.2 is an LD plot of the pairwise D' across the eight candidate genes. The magnitude of the gene sets reflects the observations pointed out in Table 4.1. The genes with a smaller number of SNPs (GNPDA2, MTCH2, SH2B1, and MC4R) revealed an overall stronger LD pattern compared to NEGR1 and FTO (genes that had more SNPs) that consisted of small blocks of LD. None of the genes appeared to be in LD with each

other, as we suspected and designed into our simulation scheme (Section 4.2.7).

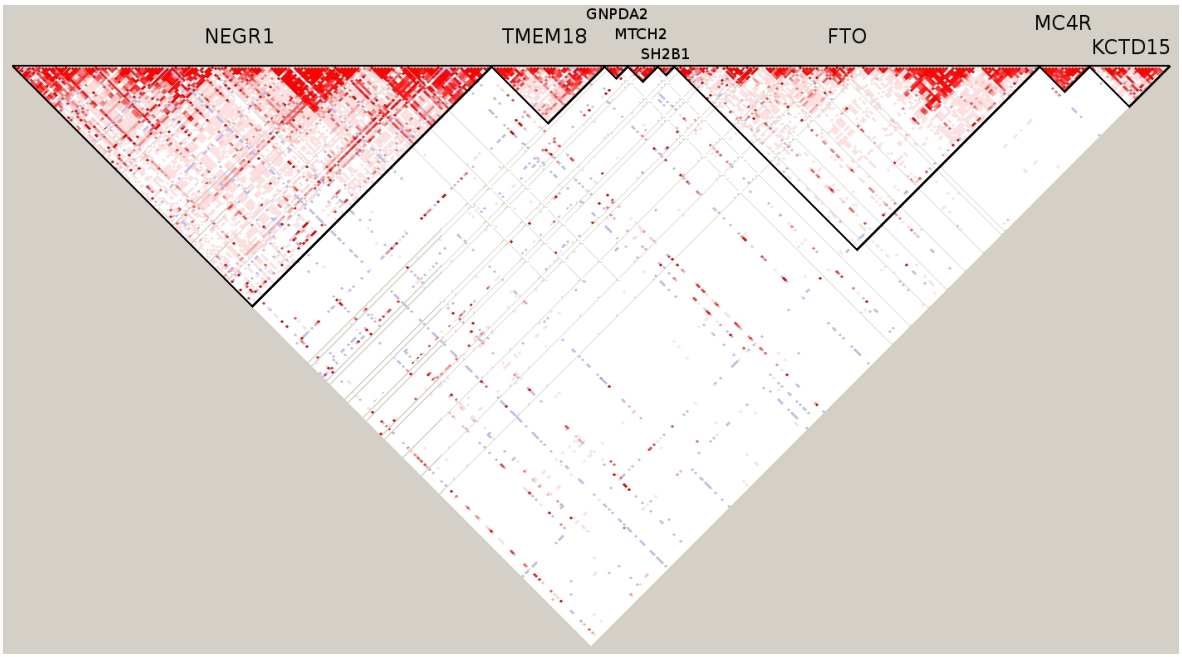

FIGURE 4.2: LD plot (D') of the eight candidate genes (NEGR1, TMEM18, GNDPA2, MTCH2, SH2B1, FTO, MC4R, and KCTD15; residing on chromosomes 1, 2, 4, 11, 16, 18, and 19; 348 representative SNPs on the Illumina HumanHap550 genotype platform; a total of 2.01 megabases in length) in the iControlDB sample of Caucasians (N = 3,172 subjects). Each gene is flanked by a set of SNPs spanning approximately 50 kilobase pairs up and downstream of the gene of interest. White represents $D' < 1$ and LOD $<$ 2, shades of pink/red $D' < 1$ and LOD ≥ 2, blue $D' = 1$ and LOD < 2, and bright red $D' = 1$ and LOD ≥ 2.

Table 4.2 presents the power of three gene-based p-value methods in analyzing the FTO gene: 1) the overall minimum p-value computed across all SNPs, permutation adjusted; 2) a consensus p-value calculated under the Bonferroni correction using the minimum p-values by gene, permutation adjusted; 3) the same gene p-values determined in method 2, though employing Fisher's Method to combine the p-values. The sections of columns in the table contain the various modeling scenarios considered. Specifically, these models were: 1) a single variant contributed to the phenotypic variation; 2) two SNPs were causal; 3) two SNPs were causal in addition to their *positive* interactive

effect; 4) two SNPs were causal in addition to their *negative* interactive effect.

The highest powers achieved for the three minimum p-value based methods was for the two SNP model that included the primary SNP (rs8050136) and the SNP that was in low LD with it (rs16953002), with all three powers at 0.90 (Table 4.2). This is likely due to the possibility of having two chances of detecting a causal variant instead of one chance, as in the case of the single SNP models. For the single SNP models, the powers were about 0.80 for the SNP 1 QTL and 0.70 for the SNP 2 QTL.

In contrast, there were significantly *lower* powers (around 0.77; Table 4.2) under the two SNP model containing SNP 1 and the SNP in modest LD with it (rs10521307). These powers were in the neighborhood of the powers under the single SNP models for SNP 1 (0.80) and SNP 3 (0.78).

Modeling interactive effects between the two SNPs resulted in similar powers for the SNP in low LD and for the SNP in modest LD with SNP 1 (around 0.90 and 0.77, respectively; Table 4.2) when specifying the *positive* interaction parameter. However, upon setting the interaction parameter with the *negative* parameter, the powers of the three methods dropped considerably (0.31 and 0.10 for the SNP 1/SNP 2 and SNP 1/SNP 3 models, respectively).

TABLE 4.2: Power of three gene-based p-value methods in analyzing the FTO gene, under single and two SNP models (with and without interactions): permutation adjusted overall minimum p-value (across all SNPs), Bonferroni adjustment on the minimum p-values stratified by gene, and Fisher's Method on the minimum p-values stratified by gene.

	Single SNP Models			Two SNP Models					
				No Interactions		+ Interactions		− Interactions	
Method	SNP 1	(SNP 2) Low LD	(SNP 3) Mod. LD	SNP 1/ Low LD	SNP 1/ Mod. LD	SNP 1/ Low LD	SNP 1/ Mod. LD	SNP 1/ Low LD	SNP 1/ Mod. LD
Overall Min P-Value	0.81	0.71	0.78	0.90	0.78	0.90	0.77	0.31	0.10
Bonferroni (Gene)	0.81	0.71	0.78	0.90	0.78	0.90	0.77	0.31	0.10
Fisher's Method (Gene)	0.79	0.70	0.78	0.90	0.76	0.89	0.76	0.30	0.10

SNP 1: rs8050136
SNP 2: rs16953002, which is in "low" LD with SNP 1 ($D' = 0.056$)
SNP 3: rs10521307, which is in "modest" LD with SNP 1 ($D' = 0.311$)
Interactions: included in the model were pairwise SNP x SNP interactions involving the two SNPs
+/− Interactions: the regression parameter of the interaction term changed the phenotypic mean in the *positive/negative* direction for every *positive* unit change in the interaction predictor
The power calculations were based on 100 simulated data sets
3,000 subjects were simulated for each data set
500 permutations were carried out for the permutation testing
The overall type I error was fixed at 0.05
For the single SNP models, R^2 was set at 0.5%, corresponding to effect sizes of 0.1021, 0.1319, and 0.1094 for SNPs 1, 2, and 3, respectively
For the two SNP models, R^2 *for each SNP* was set at 0.5%, therefore the effect sizes as specified above were utilized
For the two SNP models including interactions, R^2 *for each SNP* was set at 0.25%, corresponding to effect sizes of 0.07220, 0.09324, and 0.07732 for SNPs 1, 2, and 3, respectively
The interaction parameters explained 0.5% of the total phenotypic variation (i.e. $R^2 = 0.5\%$), corresponding to positive interaction parameters of 0.05659 and 0.04185 and negative parameters of -0.2331 and -0.1876 (SNP 1 x SNP 2 and SNP 1 x SNP 3)
The quantitative trait was randomly generated from a standard normal distribution (i.e. mean zero and standard deviation of one)

Table 4.3 presents the power results of PLINK's Set-Based Association Test (SBT) and our proposed step-wise forward SNP selection procedure ("Step") in analyzing the FTO gene under the single and two SNP causal models. We note that although we imposed the r^2 threshold for values of 1, 0.8, 0.5, and 0.1, we did not include these results in Tables 4.3 and 4.4 because it did not appear that this threshold type impacted the power to any appreciable degree. With SNP 1 (rs8050136) as the sole QTL, PLINK attained the highest power above all other methods (0.86). However, the more stringent p-value threshold (0.0025), adversely affected PLINK's power whereas it substantially improved our step-wise method (power = 0.80), though it did not greatly improve upon the three minimum p-value based methods. Under the SNP 2 single QTL model (rs16953002, in low LD with SNP 1), PLINK and our method performed almost equivalently, with the exception of p-value threshold 1 / max 10 and p-value threshold 0.05 / max 5 in which our method performed about 10% in power better. At the more stringent thresholds (p-value = 0.0025 and max = 5 or 10), both PLINK and "Step" were detecting this SNP at about the same rate as the minimum p-value based methods. Lastly, under the SNP 3 QTL model (rs10521307, in modest LD with SNP 1), both PLINK and Step had similar powers, with the exception of thresholds p-value = 0.05 and max = 10, for which we had an improvement of power of 16%. Similar to the SNP 2 QTL model, PLINK and Step almost reached the power of the minimum p-value based methods (power = 0.78; Table 4.2).

As for the two SNP models, PLINK outperformed all of the methods, including the minimum p-value based ones, at a power of 0.93 and 0.88 (Table 4.3; SNP 1/SNP 2 and SNP 1/SNP 3, respectively) versus 0.90 and 0.77 for the minimum p-value methods (Table 4.2; SNP 1/SNP 2 and SNP 1/SNP 3, respectively). However, as we observed before under the single SNP model of SNP 1, at the more rigorous p-value threshold of 0.0025, PLINK had the worse power amongst all methods (0.85 and 0.70). Under

TABLE 4.3: Power of PLINK's Set-Based Association Test (SBT) and our proposed step-wise forward SNP selection procedure ("Step") in analyzing the FTO gene, under the single and two SNP causal models.

		Single SNP Models						Two SNP Models			
Thresholds		SNP 1		(SNP 2) Low LD		(SNP 3) Mod. LD		SNP 1/ Low LD		SNP 1/ Mod. LD	
P-Value	Max	SBT	Step	SBT	Step	SBT	Step	SBT	Step	SBT	Step
1.0000	5	0.85	0.67	0.52	0.59	0.64	0.69	0.93	0.89	0.84	0.82
1.0000	10	0.86	0.59	0.42	0.53	0.59	0.62	0.91	0.83	0.88	0.82
0.0500	5	0.86	0.65	0.54	0.62	0.62	0.65	0.93	0.88	0.83	0.83
0.0500	10	0.84	0.60	0.52	0.56	0.46	0.62	0.88	0.87	0.79	0.81
0.0025	5	0.72	0.80	0.70	0.69	0.74	0.76	0.84	0.89	0.70	0.85
0.0025	10	0.72	0.80	0.68	0.69	0.77	0.76	0.85	0.89	0.70	0.85

SNP 1: rs8050136
SNP 2: rs16953002, which is in "low" LD with SNP 1 ($D' = 0.056$)
SNP 3: rs10521307, which is in "modest" LD with SNP 1 ($D' = 0.311$)
The power calculations were based on 100 simulated data sets
3,000 subjects were simulated for each data set
500 permutations were carried out for the permutation testing
The overall type I error was fixed at 0.05
For the single SNP models, R^2 was set at 0.5%, corresponding to effect sizes of 0.1021, 0.1319, and 0.1094 for SNPs 1, 2, and 3, respectively
For the two SNP models, R^2 *for each SNP* was set at 0.5%, therefore the effect sizes as specified above were utilized
The quantitative trait was randomly generated from a standard normal distribution (i.e. mean zero and standard deviation of one)

the two SNP model that included the SNP in moderate LD with SNP 1, there was a much greater gain in power utilizing our step-wise approach as compared to PLINK's SBT (SBT = 0.70 and Step = 0.85). In fact, our step-wise approach at this p-value threshold of 0.0025 had greater power than the minimum p-value approaches.

Table 4.4 contains the power of the SBT and Step in analyzing the FTO candidate gene, under the two SNP causal models including interactions. For the models with the SNP 1/SNP 2 positive interactions, both the SBT and Step (around 0.92) had powers slightly better than the three minimum p-value based methods (0.90; Table 4.2). However, for the SNP 1/SNP 3 positive interactions, both SBT and Step were

more powerful (0.87 and 0.83, respectively) than the minimum p-value methods (0.77). In some threshold instances, the SBT did better than Step and vice versa. Lastly, under the negative interaction models, Step was the most powerful amongst all methods (0.38 for the SNP 1/SNP 2 scenario versus 0.31 for the min p-value methods; 0.23 for the SNP 1/SNP 3 scenario versus 0.10 for the min p-value methods). Furthermore, Step seemed to consistently outperform SBT when modeling negative interactions. We note that although Step was most powerful, these powers were relatively low and surprisingly not an overall powerful method in discerning amongst the significant genetic parameters, at least for the negative interaction models.

Table 4.5 contains the results of the candidate gene list approach in which we analyzed "super" data sets (under two alternative models) comprising eight candidate gene regions. The minimum p-value based approaches assessed at the gene level suffered a substantial drop in power (less so for Fisher's Method) under the alternative model that housed the secondary QTLs in gene regions (NEGR1 and TMEM18) comprising a larger set of SNPs (Model 2) compared to Model 1 that modeled secondary QTLs in gene regions (MTCH2 and SH2B1) with less SNPs. Interestingly, computing the minimum p-value across all available SNPs in the data set increased in power for Model 2. It is notable that Fisher's Method under Model 1 performed the best (power = 0.99), which was somewhat not surprising since the gene regions bearing the simulated QTLs exhibited little to no LD (Figure 4.2), and Fisher's Method strictly assumes mutually independent tests. The Bonferroni correction also had decent power (0.93), though following Fisher's Method in robustness was our step-wise procedure at 0.96. Power marginally dropped upon imposing a stricter p-value threshold of 0.0025, which was the opposite trend of what we observed for the candidate FTO gene alternative models. PLINK showed adequate power when no p-value filter was implemented (0.86), however, its power fell dramatically for the more stringent p-value threshold (0.67).

TABLE 4.4: Power of PLINK's Set-Based Association Test (SBT) and our proposed step-wise forward SNP selection procedure ("Step") in analyzing the FTO gene, under the two SNP causal models including interactions.

		Two SNP Models							
		+ Interactions				− Interactions			
Thresholds		SNP 1/ Low LD		SNP 1/ Mod. LD		SNP 1/ Low LD		SNP 1/ Mod. LD	
P-Value	Max	SBT	Step	SBT	Step	SBT	Step	SBT	Step
1.0000	5	0.91	0.91	0.87	0.82	0.28	0.35	0.080	0.20
1.0000	10	0.88	0.76	0.83	0.60	0.23	0.38	0.092	0.23
0.0500	5	0.90	0.89	0.82	0.78	0.25	0.28	0.080	0.14
0.0500	10	0.83	0.86	0.73	0.73	0.22	0.24	0.103	0.16
0.0025	5	0.88	0.92	0.70	0.83	0.30	0.31	0.080	0.11
0.0025	10	0.87	0.92	0.70	0.83	0.31	0.31	0.080	0.11

SNP 1: rs8050136
SNP 2: rs16953002, in "low" LD with SNP 1 ($D' = 0.056$)
SNP 3: rs10521307, in "modest" LD with SNP 1 ($D' = 0.311$)
Interactions: included in the model were pairwise SNP x SNP interactions involving the two SNPs
+/− Interactions: the regression parameter of the interaction term changed the phenotypic mean in the *positive/negative* direction for every *positive* unit change in the interaction predictor
The power calculations were based on 100 simulated data sets
3,000 subjects were simulated for each data set
500 permutations were carried out for the permutation testing
The overall type I error was fixed at 0.05
The R^2 *for each SNP* was set at 0.25%, corresponding to effect sizes of 0.07220, 0.09324, and 0.07732 for SNPs 1, 2, and 3
The interaction parameters explained 0.5% of the total phenotypic variation (i.e. $R^2 = 0.5\%$), corresponding to positive interaction parameters of 0.05659 and 0.04185 and negative parameters of -0.2331 and -0.1876 (SNP 1 x SNP 2 and SNP 1 x SNP 3)
The quantitative trait was randomly generated from a standard normal distribution (i.e. mean zero and standard deviation of one)

Power remained roughly the same under Model 2 for the SBT and Step. We note that the Step attained the highest power amongst all methods under Model 2 (0.96). Overall, our step-wise method achieved substantially greater power than PLINK.

4.4 Discussion

We have reported results on some available methods for analyzing candidate gene or a set of candidate genes. In contrast to tests of association that are conducted at the SNP level, the approaches we have explored offer a consensus estimate of significance for the entire set of SNPs comprising the gene or genes. With regard to analyzing a data set consisting of multiple candidate genes, although we framed our power analyses about a list of obesity-related genes, one could plausibly apply the foregoing analytical methods to a putative biological pathway comprising various genes.

We assessed the power of three methods based on the minimum p-value for a given candidate gene or genes: 1) the overall minimum p-value computed across all SNPs, permutation adjusted; 2) a consensus p-value calculated under the Bonferroni correction using the minimum p-values by gene, permutation adjusted; 3) the same gene p-values determined in method 2, though employing Fisher's Method to combine the p-values.

In the candidate gene approach, for the simulated data sets in which we modeled an underlying single quantitative trait locus, these minimum p-value based methods had decent power for the SNP 1 (rs8050136) and SNP 3 (rs10521307) models (power = $\sim$0.80; Table 4.2) though not for the SNP 2 (rs16953002) model. Consequently, SNPs 1 and 3 had much more abundant minor allele frequencies than SNP 2 (SNP 1 MAF = 0.3989 and SNP 3 MAF = 0.2977 in comparison to SNP 2 MAF = 0.1741). Perhaps due to the less frequent presence of SNP 2's quantitative trait variant, the overall phenotypic distribution was weighted heavier with trait values from homozygotes for

TABLE 4.5: Power of the minimum p-value based methods (Overall, Bonferroni, and Fisher's Method), PLINK's Set-Based Association Test (SBT), and our proposed stepwise forward SNP selection procedure ("Step") in analyzing the eight obesity-related candidate genes in a single analysis, under two alternative modeling scenarios.

Method		Model 1[a]		Model 2[b]	
Overall Min P-Value		0.80		0.87	
Bonferroni (Gene)		0.93		0.78	
Fisher's Method (Gene)		0.99		0.93	
Thresholds		Model 1[a]		Model 2[b]	
P-Value	Max	SBT	Step	SBT	Step
1.0000	5	0.86	0.96	0.87	0.96
1.0000	10	0.81	0.94	0.87	0.96
0.0500	5	0.88	0.96	0.88	0.96
0.0500	10	0.83	0.93	0.89	0.94
0.0025	5	0.72	0.90	0.72	0.90
0.0025	10	0.67	0.90	0.60	0.90

[a] Model 1: two QTLs in FTO (rs8050136: R^2 = 0.35%/slope = 0.08543; rs10521307: R^2 = 0.25%/slope = 0.07732), one near MC4R (rs10871777: R^2 = 0.35%/slope = 0.09836), and one each in MTCH2 (rs10838738: R^2 = 0.25%/slope = 0.07460) and SH2B1 (rs7498665: R^2 = 0.25%/slope = 0.07353)

[b] Model 2: the same SNPs and effect sizes in FTO and near MC4R (noted above), and one QTL each in NEGR1 (rs2568958: R^2 = 0.25%/slope = 0.07501) and TMEM18 (rs4854344: R^2 = 0.25%/slope = 0.09115)

The power calculations were based on 100 simulated data sets

3,000 subjects were simulated for each data set

500 permutations were carried out for the permutation testing

The overall type I error was fixed at 0.05

For the minimum p-value gene-based approaches (Bonferroni and Fisher's Method), the gene boundaries were defined as the SNPs that flanked the region comprising the gene plus 50 kilobase pairs up and downstream of the gene

The quantitative trait was randomly generated from a standard normal distribution (i.e. mean zero and standard deviation of one)

the major allele. Hence, there was not such a clear distinction amongst this homozygous distribution and the distributions arising from the heterozygotes and the homozygotes for the minor allele.

As for the two SNP model results, we observed excellent power (0.90; Table 4.2) when modeling SNP 1 and SNP 2 (the SNP in low LD with SNP 1) though we observed a 12% decline in power for the SNP 1/SNP 3 models. This result can be explained in the following way. We calculated the D' between the SNPs as 0.311 whereas the r^2 was fairly low (0.02823). In addition, the minor allele frequencies of the two SNPs were different (SNP 1, rs8050136: MAF = 0.3989; SNP 3, rs10521307: MAF = 0.2977), therefore given the mathematical nature of D' and r^2, it is not surprising that their values were discordant since D' is affected by allele frequencies. These D' and r^2 results imply no (or minimal) historical recombination events such that the representative pool of haplotypes consisted of three possible haplotypes rather than all four. Taken together with the power results, this suggests an overabundance of the haplotype bearing *both* risk variants. Thus, including an additional causal SNP in simulating the quantitative trait did not improve our chances in detecting these QTLs.

With respect to the two SNP models including interactions, overall the minimum gene p-value based methods performed poorly, with the exception of the powers estimated under the SNP 1/ SNP 2 (SNP in low LD with SNP 1) interaction models. This was expected since taking the minimum gene p-value completely ignored the effects of interaction on the quantitative trait.

Under the single and two SNP alternative models without interactions, PLINK performed fairly well in comparison to our step-wise procedure. Taken together with our observation that pruning SNPs in high LD with the significant SNPs did not positively nor adversely affect power, PLINK's approach of averaging the test statistics did not push the averaged test statistic toward non-significance but rather this provided more

power. This was counter-intuitive to us as we expected the average test statistic to be less in magnitude than the overall minimum p-value across all SNPs.

On the contrary, under the single and two SNP alternative models without interactions, our step-wise approach in many instances had significantly greater power than PLINK when restricting individual p-values to be less than 0.0025. However, under these alternative models and individual p-value restriction, we were not able to exceed the power attained by PLINK. A possible explanation for this drop in power with PLINK's method in these instances is the following. Upon removing SNPs that did not meet the 0.0025 p-value threshold, the SNP or SNPs that drove the signal in the averaged test statistic had corresponding p-values above 0.0025. Thus, in assessing statistical significance of the observed average test statistic via permutation testing, it was quite probable to calculate an average test statistic that was as extreme or more extreme than the observed. On the other hand, for the instances in which power increased under the 0.0025 p-value filter, the SNP or SNPs largely responsible for the signal were likely retained after applying the filter.

As for the candidate gene list analyses, the minimum p-value approaches showed acceptable power, though there were clearly gains in power in using our step-wise method. Compared to the Bonferroni adjustment, Fisher's method was more powerful under both alternative model scenarios. Clearly, there was less of a penalty (or perhaps no penalty at all) for multiple testing in combining the gene-specific permutation adjusted minimum p-values using Fisher's Method than Bonferroni. Regarding the almost perfect power of Fisher's Method under Model 1 (0.99), there were probably strong associations at each gene, coupled with the small penalty for combining the p-values and that each permutation test was for all practical purposes independent (due to the minimal LD amongst the gene regions shown in Figure 4.2). The significant decrease in power observed under Model 2 for the minimum p-value based methods at

the gene level (not so much for Fisher's Method) was likely due to the fact that two of the QTL signals (in NEGR1 and TMEM18) were embedded amongst a greater number of SNPs in contrary to MTCH2 and SH2B1 (Model 1). Fundamentally, under Model 2 there was more noise introduced that the methods had to tease out.

PLINK's approach of averaging the test statistics did not prove to be much better than the minimum p-value based methods, for the candidate gene list analyses. Though the advantage to PLINK's technique is that the analyst is not required to *a priori* specify each gene region. This offers a more unbiased approach in computing a consensus p-value for a candidate gene list of biological pathway. This was also a desirable feature for our step-wise method.

In terms of the candidate gene list analyses, our step-wise forward SNP selection procedure built on a linear regression modeling framework proved to be a more powerful approach than PLINK. A possible explanation is the following. Under both modeling scenarios, two SNPs contributed more to the phenotypic variation than the other set of three SNPs ($R^2 = 0.35\%$ for the primary two QTLs and $R^2 = 0.25\%$ for the secondary three QTLs). Since PLINK averages the test statistics in order to offer a proxy p-value, the stronger associations from the primary QTLs were likely dampened by the associations from the secondary QTLs, upon computing the average. On the contrary, our step-wise method jointly tested the association at the various loci and was not greatly disadvantaged by a mixture of strong and weaker QTLs.

Although we discussed (Section 4.2) and implemented an alternative stopping criterion for our step-wise forward SNP selection technique (which was less stringent in including a putative SNP in the SNP covariate set), we did not present these results because the powers estimated were strikingly similar to the more rigorous approach of ceasing to build the SNP covariate set at the first joint p-value that did not improve upon the prior. In addition, during exploratory analyses we also implemented

a third stopping criterion in which we mimicked PLINK's set-based test criteria. In other words, as long as a SNP fulfilled the r^2 and individual p-value filters and the set size did not exceed the maximum specified, the SNP in question was entered into the set. The power results based on this third stopping criterion also were quite similar to the first rule presented in the Tables, therefore we did not consider this rule in further analyses.

Furthermore, pruning SNPs in "high" LD with significant SNPs also did not appear to affect the powers of PLINK's test nor ours, and so we presented the results for which no LD pruning filter was turned on. This result suggests that SNPs in LD with significant SNPs (as determined by PLINK's SBT or our step-wise procedures) did not add noise nor did they necessarily positively contribute in detecting a true association with the quantitative trait.

In summary, we have presented results of methods that offer a consensus p-value for analyzing data from a candidate gene study, or a study involving a list of candidate genes or biological pathway. Each method bears advantages and disadvantages, and under certain scenarios some methods performed better than others. There was no single method that proved to be the best across all modeling and study design scenarios.

CHAPTER 5

NATURAL EXTENSIONS TO THE CURRENT INVESTIGATIONS

5.1 Using Public Control Genotype Data to Increase Power and Decrease Cost of Case-Control Genetic Association Studies

There has been an increased use of imputing unobserved SNPs across commercial genotyping platforms, which has the advantage of enabling researchers to more exhaustively test for association on a denser set of SNPs, by capitalizing on the available LD information contained across the genotyped loci. In terms of our proposed two-stage replication-based design, further investigations could assess the impacts on power and type I error of imputing untyped SNPs in the genome-wide association phase of the study (i.e. stage one). Because the study cases and public controls were genotyped at different points in time and also perhaps on different platforms, the question arises

on how best to deal with possible batch effects in terms of imputing SNPs. A clear advantage of imputing SNPs in stage one is that the study cases and public controls need not be initially genotyped on the same platform. However, to what extent is the type I error affected by this and how much gain in power would imputation offer?

In addition, rare SNPs would presumably be imputed with a greater degree of uncertainty as compared to common SNPs. One could investigate overall power as a function of this uncertainty, over a range of minor allele frequencies (e.g. from incredibly rare such as $< 1\%$ to more common such as 30%).

Granted that an imputed SNP in stage one is selected for follow-up in stage two, how does this also influence power of our proposed two-stage design? For example, as the proportion of true findings at imputed loci increases, what is the trend in power?

As a last point regarding imputation, how do we best handle SNPs with low to high quality scores? As an extreme scenario, how powerful is our two-stage design for low quality SNPs that truly are disease causing agents? Could we devise an analytical approach that would account for such poor quality SNPs or even improve upon their reliability somehow?

Finally, it would be fascinating to design and implement a study in practice that utilizes our proposed two-stage replication-based study design. Such empirical results could offer invaluable insight to the utility of our design in a practical setting, as compared to our hypothesized alternative disease models.

5.2 Haplotype Sharing Methods in Association Studies

In comparing our haplotype sharing methods with currently available techniques, we assumed that phase was known. This may or may not be a reasonable assumption

in certain settings, thus some possibilities could be to explore alternatives to this assumption. For example, we could develop sharing statistics in which phase is unknown, therefore we utilize comparisons in genotypes rather than haplotypes. The sharing measures would have to incorporate the phase ambiguity such as the use of penalties or determining weights for the most probable pairs of alleles within genotypes, built upon the extent of sharing observed throughout the sample.

If we impose phase as a preliminary pre-processing step, then we could examine alternative ways to phase the genotype data, other than fastPHASE, the method we employed for our study. For example, several programs utilize the expectation-maximization (EM) algorithm in estimating haplotype signatures, such as PLINK, Haploview, and haplo.stats. In addition, SimWalk2 uses a type of simulated annealing approach to phase haplotypes. Would the haplotype measures assessed in our study perform the same, better, or worse if we employed other methods to infer haplotype phase?

5.3 Gene and Pathway-Based P-Values

An alternative to the proposed step-wise forward SNP selection method (Section 4.2) would be a haplotype-based approach instead of a genotype-based approach to facilitate the simultaneous analysis of multiple SNPs within a candidate gene. For example, one could carry out a conventional haplotype χ^2 test on the observed set of unique haplotype signatures by categorizing the haplotypes on case-control status, as described in Section 3.2.7. In this way, all unique haplotype signatures would be assumed to be ancestrally distinct, although consequently the formal test could have many degrees of freedom. Optionally, the analyst could a priori classify the observed unique haplotype signatures into groups if the plausibility exists that subgroups of the haplotypes originated from

the same lineage or could allow the data to cluster individual haplotypes into different "risk groups". This would invariably decrease the degrees of freedom of the test and increase power to detect an association, though allowing the data to generate haplotype groupings would require a permutation test to account for the data driven nature of the clustering.

We note that the aforementioned haplotype-based approach readily facilitate the inclusion of *a priori* hypotheses (for example, the combining of putatively functional alleles or genotypes based on public data bases) that, if properly applied, can increase power. The potential advantage of such an approach should be its flexibility in allowing the data to determine the optimal SNPs/haplotypes to include in the final test statistics, though this greater flexibility could actually result in decreased power if good *a priori* information is available and not incorporated or if results can in large be explained by a single SNP.

In our current investigation, we explored a limited example of alternative models that included SNP x SNP interactions, as well as one example of a list of candidate genes. We could further our work by simulating larger or smaller lists of genes as well as a putative biological pathway instead of a simple list of genes taken from a prior study. Lastly, due to the computationally intensive nature of our step-wise forward SNP selection procedure in which we employed parallel computing on high-end super-computers, we could develop ways that are less computational such that our method could be employed on a genome-wide scale.

REFERENCES

Ahn, K., Haynes, C., Kim, W., Fleur, R. S., Gordon, D. and Finch, S. J. (2007). The effects of SNP genotyping errors on the power of the Cochran-Armitage linear trend test for case/control association studies. *Ann Hum Genet* **71**, 249–261.

Akey, J., Jin, L. and Xiong, M. (2001). Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet* **9**, 291–300.

Altshuler, D., Hirschhorn, J. N., Klannemark, M., Lindgren, C. M., Vohl, M. C., Nemesh, J., Lane, C. R., Schaffner, S. F., Bolk, S., Brewer, C., Tuomi, T., Gaudet, D., Hudson, T. J., Daly, M., Groop, L. and Lander, E. S. (2000). The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* **26**, 76–80.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386.

Bader, J. S. (2001). The relative power of SNPs and haplotype as genetic markers for association tests. *Pharmacogenomics* **2**, 11–24.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7**, 781–791.

Barrett, J. C. and Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nat Genet* **38**, 659–662.

Barrett, J. C., Fry, B., Maller, J. and Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265.

Beckmann, L., Thomas, D. C., Fischer, C. and Chang-Claude, J. (2005). Haplotype sharing analysis using mantel statistics. *Hum Hered* **59**, 67–78.

Berger, M., Stassen, H. H., Khler, K., Krane, V., Mnks, D., Wanner, C., Hoffmann, K., Hoffmann, M. M., Zimmer, M., Bickebller, H. and Lindner, T. H. (2006). Hidden population substructures in an apparently homogeneous population bias association studies. *Eur J Hum Genet* **14**, 236–244.

Bertina, R. M., Koeleman, B. P., Koster, T., Rosendaal, F. R., Dirven, R. J., de Ronde, H., van der Velden, P. A. and Reitsma, P. H. (1994). Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* **369**, 64–67.

Boehnke, M. (1994). Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. *Am J Hum Genet* **55**, 379–390.

Boehnke, M. and Langefeld, C. D. (1998). Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *Am J Hum Genet* **62**, 950–961.

Borecki, I. B. and Suarez, B. K. (2001). Linkage and association: basic concepts. *Adv Genet* **42**, 45–66.

Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33 Suppl**, 228–237.

Bourgain, C., Genin, E., Quesneville, H. and Clerget-Darpoux, F. (2000). Search for multifactorial disease susceptibility genes in founder populations. *Ann Hum Genet* **64**, 255–265.

Bukszár, J. and van den Oord, E. J. C. G. (2006a). Accurate and efficient power calculations for 2 x m tables in unmatched case-control designs. *Stat Med* **25**, 2632–2646.

Bukszár, J. and van den Oord, E. J. C. G. (2006b). Optimization of two-stage genetic designs where data are combined using an accurate and efficient approximation for Pearson's statistic. *Biometrics* **62**, 1132–1137.

Burgtorf, C., Kepper, P., Hoehe, M., Schmitt, C., Reinhardt, R., Lehrach, H. and Sauer, S. (2003). Clone-based systematic haplotyping (CSH): a procedure for physical haplotyping of whole genomes. *Genome Res* **13**, 2717–2724.

Cardon, L. R. and Bell, J. I. (2001). Association study designs for complex diseases. *Nat Rev Genet* **2**, 91–99.

Cardon, L. R. and Palmer, L. J. (2003). Population stratification and spurious allelic association. *Lancet* **361**, 598–604.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q. and Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* **22**, 231–238.

Carlson, C. S., Eberle, M. A., Kruglyak, L. and Nickerson, D. A. (2004). Mapping complex disease loci in whole-genome association studies. *Nature* **429**, 446–452.

Chakravarti, A. (1999). Population genetics–making sense out of sequence. *Nat Genet* **21**, 56–60.

Chapman, D. G. and Nam, J. M. (1968). Asymptotic power of chi square tests for linear trends in proportions. *Biometrics* **24**, 315–327.

Chapman, J. and Whittaker, J. (2008). Analysis of multiple SNPs in a candidate gene or region. *Genet Epidemiol* **32**, 560–566.

Chapman, J. M., Cooper, J. D., Todd, J. A. and Clayton, D. G. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* **56**, 18–31.

Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L., Buchanan, A., Stengrd, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E. and Sing, C. F. (1998). Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet* **63**, 595–612.

Clayton, D. (1999). A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet* **65**, 1170–1177.

Clayton, D., Chapman, J. and Cooper, J. (2004). Use of unphased multilocus genotype data in indirect association studies. *Genet Epidemiol* **27**, 415–428.

Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C., Ovington, N. R., Stevens, H. E., Nutland, S., Howson, J. M. M., Faham, M., Moorhead, M., Jones, H. B., Falkowski, M., Hardenbol, P., Willis, T. D. and Todd, J. A. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* **37**, 1243–1246.

Cochran, W. G. (1954). Some methods for strengthening the common chi-squared tests. *Biometrics* **10**, 417–451.

Collins, F. S., Guyer, M. S. and Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581.

Complex Trait Consortium (2003). The nature and identification of quantitative trait loci: a community's view. *Nat Rev Genet* **4**, 911–916.

Concannon, P., Gogolin-Ewens, K. J., Hinds, D. A., Wapelhorst, B., Morrison, V. A., Stirling, B., Mitra, M., Farmer, J., Williams, S. R., Cox, N. J., Bell, G. I., Risch, N. and Spielman, R. S. (1998). A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. *Nat Genet* **19**, 292–296.

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., Roses, A. D., Haines, J. L. and Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921–923.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nat Genet* **29**, 229–232.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*. John Wiley and Sons, New York, second edition.

Drysdale, C. M., McGraw, D. W., Stack, C. B., Stephens, J. C., Judson, R. S., Nandabalan, K., Arnold, K., Ruano, G. and Liggett, S. B. (2000). Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc Natl Acad Sci U S A* **97**, 10483–10488.

Edwards, B. J., Haynes, C., Levenstien, M. A., Finch, S. J. and Gordon, D. (2005). Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. *BMC Genet* **6**, 18.

Ellsworth, D. L. and Manolio, T. A. (1999a). The emerging importance of genetics in epidemiologic research. I. Basic concepts in human genetics and laboratory technology. *Ann Epidemiol* **9**, 1–16.

Ellsworth, D. L. and Manolio, T. A. (1999b). The emerging importance of genetics in epidemiologic research II. Issues in study design and gene mapping. *Ann Epidemiol* **9**, 75–90.

Ellsworth, D. L. and Manolio, T. A. (1999c). The emerging importance of genetics in epidemiologic research III. Bioinformatics and statistical genetic methods. *Ann Epidemiol* **9**, 207–224.

Elston, R. C., Buxbaum, S., Jacobs, K. B. and Olson, J. M. (2000). Haseman and Elston revisited. *Genet Epidemiol* **19**, 1–17.

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* **12**, 921–927.

Falk, C. T. and Rubinstein, P. (1987). Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* **51**, 227–233.

Fan, R. and Knapp, M. (2003). Genome association studies of complex diseases by case-control designs. *Am J Hum Genet* **72**, 850–868.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. Oliver and Boyd, London, 4 edition.

Forrest, W. F. (2001). Weighting improves the "new Haseman-Elston" method. *Hum Hered* **52**, 47–54.

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R. B., Elliott, K. S., Lango, H., Rayner, N. W., Shields, B., Harries, L. W., Barrett, J. C., Ellard, S., Groves, C. J., Knight, B. et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to

childhood and adult obesity. *Science* **316**, 889–894.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J. and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229.

Grant, S. F. A., Li, M., Bradfield, J. P., Kim, C. E., Annaiah, K., Santa, E., Glessner, J. T., Casalunovo, T., Frackelton, E. C., Otieno, F. G., Shaner, J. L., Smith, R. M., Imielinski, M., Eckert, A. W., Chiavacci, R. M., Berkowitz, R. I. and Hakonarson, H. (2008). Association analysis of the FTO gene with obesity in children of Caucasian and African ancestry reveals a common tagging SNP. *PLoS One* **3**, e1746.

Gruber, M. H. J. (1998). *Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators*. Marcel Dekker Inc., New York.

Haldane, J. B. S. and Smith, C. A. B. (1947). A new estimate of the linkage between the genes for colourblindness and haemophilia in man. *Ann Eugen* **14**, 10–31.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**, 95–108.

Hollox, E. J., Poulter, M., Zvarik, M., Ferak, V., Krause, A., Jenkins, T., Saha, N., Kozlov, A. I. and Swallow, D. M. (2001). Lactase haplotype diversity in the Old World. *Am J Hum Genet* **68**, 160–172.

Hom, G., Graham, R. R., Modrek, B., Taylor, K. E., Ortmann, W., Garnier, S., Lee, A. T., Chung, S. A., Ferreira, R. C., Pant, P. V. K., Ballinger, D. G., Kosoy, R. et al. (2008). Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. *N Engl J Med* **358**, 900–909.

Horvath, S. and Laird, N. M. (1998). A discordant-sibship test for disequilibrium and linkage: no need for parental data. *Am J Hum Genet* **63**, 1886–1897.

International HapMap Consortium (2003). The International HapMap Project. *Nature* **426**, 789–796.

International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**, 1299–1320.

Jeffreys, A. J., Kauppi, L. and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* **29**, 217–222.

Jimenez-Sanchez, G., Childs, B. and Valle, D. (2001). Human disease genes. *Nature* **409**, 853–855.

Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J. and de Bakker, P. I. W. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–2939.

Kang, H., Qin, Z. S., Niu, T. and Liu, J. S. (2004). Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphisms. *Am J Hum Genet* **74**, 495–510.

Kooperberg, C. and Leblanc, M. (2008). Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet Epidemiol* **32**, 255–263.

Kruglyak, L. and Nickerson, D. A. (2001). Variation is the spice of life. *Nat Genet* **27**, 234–236.

Laan, M. and Pbo, S. (1997). Demographic history and linkage disequilibrium in human populations. *Nat Genet* **17**, 435–438.

Lander, E. S. (1996). The new genomics: global views of biology. *Science* **274**, 536–539.

Lander, E. S. and Schork, N. J. (1994). Genetic dissection of complex traits. *Science* **265**, 2037–2048.

Lange, E. M. and Boehnke, M. (2004). The haplotype runs test: the parent-parent-affected offspring trio design. *Genet Epidemiol* **27**, 118–130.

Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83**, 311–321.

Li, M., Wang, K., Grant, S. F. A., Hakonarson, H. and Li, C. (2009). ATOM: a powerful gene-based association test by combining optimally weighted markers. *Bioinformatics* **25**, 497–503.

Long, A. D. and Langley, C. H. (1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res* **9**, 720–731.

Luca, D., Ringquist, S., Klei, L., Lee, A. B., Gieger, C., Wichmann, H.-E., Schreiber, S., Krawczak, M., Lu, Y., Styche, A., Devlin, B., Roeder, K. and Trucco, M. (2008). On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *Am J Hum Genet* **82**, 453–463.

Malo, N., Libiger, O. and Schork, N. J. (2008). Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am J Hum Genet* **82**, 375–385.

Maniatis, N., Collins, A., Xu, C. F., McCarthy, L. C., Hewett, D. R., Tapper, W., Ennis, S., Ke, X. and Morton, N. E. (2002). The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc Natl Acad Sci U S A* **99**, 2228–2233.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res* **27**, 209–220.

Marchini, J., Cardon, L. R., Phillips, M. S. and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nat Genet* **36**, 512–517.

Matise, T. C., Chen, F., Chen, W., Vega, F. M. D. L., Hansen, M., He, C., Hyland, F. C. L., Kennedy, G. C., Kong, X., Murray, S. S., Ziegle, J. S., Stewart, W. C. L. and Buyske, S. (2007). A second-generation combined linkage physical map of the human genome. *Genome Res* **17**, 1783–1786.

McDonald, O. G., Krynetski, E. Y. and Evans, W. E. (2002). Molecular haplotyping of genomic DNA for multiple single-nucleotide polymorphisms located kilobases apart using long-range polymerase chain reaction and intramolecular ligation. *Pharmacogenetics* **12**, 93–99.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157.

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R. and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584.

Mitchell, A. A., Cutler, D. J. and Chakravarti, A. (2003). Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am J Hum Genet* **72**, 598–610.

Morris, R. W. and Kaplan, N. L. (2002). On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol* **23**, 221–233.

Morton, N. E. and Collins, A. (1998). Tests and estimates of allelic association in complex inheritance. *Proc Natl Acad Sci U S A* **95**, 11389–11393.

Moskvina, V., Craddock, N., Holmans, P., Owen, M. J. and O'Donovan, M. C. (2006). Effects of differential genotyping error rate on the type I error probability of case-control studies. *Hum Hered* **61**, 55–64.

Moskvina, V., Holmans, P., Schmidt, K. M. and Craddock, N. (2005). Design of case-controls studies with unscreened controls. *Ann Hum Genet* **69**, 566–576.

National Cancer Institute (2009a). PDQ cancer information summary, genetics of breast

and ovarian cancer - health professional. Technical report, National Cancer Institute, Bethesda, MD.

National Cancer Institute (2009b). SEER cancer statistics review, 1975–2005. Technical report, National Cancer Institute, Bethesda, MD.

National Center for Biotechnology Information, Entrez Gene (2009). PHB prohibitin [Homo sapiens].

Neale, B. M. and Purcell, S. (2008). The positives, protocols, and perils of genome-wide association. *Am J Med Genet B Neuropsychiatr Genet* **147B**, 1288–1294.

Niu, T. (2004). Algorithms for inferring haplotypes. *Genet Epidemiol* **27**, 334–347.

Nolte, I. M. (2002). *Statistics and population genetics of haplotype sharing as a tool for fine-mapping of disease gene loci*. PhD thesis, University of Groningen.

Nolte, I. M., de Vries, A. R., Spijker, G. T., Jansen, R. C., Brinza, D., Zelikovsky, A. and te Meerman, G. J. (2007). Association testing by haplotype-sharing methods applicable to whole-genome analysis. *BMC Proc* **1 Suppl 1**, S129.

Nolte, I. M. and te Meerman, G. J. (2002). The probability that similar haplotypes are identical by descent. *Ann Hum Genet* **66**, 195–209.

Odeberg, J., Holmberg, K., Eriksson, P. and Uhln, M. (2002). Molecular haplotyping by pyrosequencing. *Biotechniques* **33**, 1104, 1106, 1108.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P. and Cox, D. R. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723.

Pecioska, S., Zillikens, M. C., Henneman, P., Snijders, P. J., Oostra, B. A., van Duijn, C. M. and Aulchenko, Y. S. (2010). Association between type 2 diabetes loci and measures of fatness. *PLoS One* **5**, e8541.

Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., Zhou, X., Reveille, J. D., Jin, L., Amos, C. I. and Xiong, M. (2009). Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur J Hum Genet* .

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (2002). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 2nd edition.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and

Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–909.

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69**, 124–137.

Purcell, S. (2007). PLINK v0.99s, http://pngu.mgh.harvard.edu/purcell/plink/.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J. and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reich, D. E. and Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends Genet* **17**, 502–510.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856.

Roeder, K. and Luca, D. (2009). Searching for disease susceptibility variants in structured populations. *Genomics* **93**, 1–4.

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261.

Satagopan, J. M., Venkatraman, E. S. and Begg, C. B. (2004). Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* **60**, 589–597.

Satagopan, J. M., Verbel, D. A., Venkatraman, E. S., Offit, K. E. and Begg, C. B. (2002). Two-stage designs for gene-disease association studies. *Biometrics* **58**, 163–170.

Schaid, D. J. (2004a). Evaluating associations of haplotypes with traits. *Genet Epidemiol* **27**, 348–364.

Schaid, D. J. (2004b). Genetic epidemiology and haplotypes. *Genet Epidemiol* **27**, 317–320.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. and Poland, G. A.

(2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* **70**, 425–434.

Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**, 629–644.

Schork, N. and Chakravarti, A. (1996). *Molecular Genetics and Gene Therapy of Cardiovascular Disease*, chapter 2: A nonmathematical overview of modern gene mapping techniques applied to human diseases, pages 79–109. Marcel Dekker Inc., New York, NY.

Schork, N. J., Fallin, D., Thiel, B., Xu, X., Broeckel, U., Jacob, H. J. and Cohen, D. (2001). The future of genetic case-control studies. *Adv Genet* **42**, 191–212.

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., Erdos, M. R., Stringham, H. M., Chines, P. S., Jackson, A. U., Prokunina-Olsson, L., Ding, C. J., Swift, A. J., Narisu, N., Hu, T., Pruim, R. et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345.

Scuteri, A., Sanna, S., Chen, W.-M., Uda, M., Albai, G., Strait, J., Najjar, S., Nagaraja, R., Orr, M., Usala, G., Dei, M., Lai, S., Maschio, A., Busonero, F., Mulas, A., Ehret, G. B., Fink, A. A., Weder, A. B., Cooper, R. S., Galan, P., Chakravarti, A., Schlessinger, D., Cao, A., Lakatta, E. and Abecasis, G. R. (2007). Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet* **3**, e115.

Sham, P. (1998). *Statistics in Human Genetics*. Hodder Arnold, London.

Silverberg, M. S., Cho, J. H., Rioux, J. D., McGovern, D. P. B., Wu, J., Annese, V., Achkar, J.-P., Goyette, P., Scott, R., Xu, W., Barmada, M. M., Klei, L., Daly, M. J. et al. (2009). Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat Genet* **41**, 216–220.

Skol, A. D., Scott, L. J., Abecasis, G. R. and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* **38**, 209–213.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T. J., Montpetit, A., Pshezhetsky, A. V., Prentki, M., Posner, B. I., Balding, D. J., Meyre, D., Polychronakos, C. and Froguel, P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885.

Slager, S. L., Huang, J. and Vieland, V. J. (2000). Effect of allelic heterogeneity on the

power of the transmission disequilibrium test. *Genet Epidemiol* **18**, 143–156.

Slager, S. L. and Schaid, D. J. (2001). Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. *Hum Hered* **52**, 149–153.

Spielman, R. S. and Ewens, W. J. (1996). The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* **59**, 983–989.

Spielman, R. S. and Ewens, W. J. (1998). A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* **62**, 450–458.

Spielman, R. S., McGinnis, R. E. and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**, 506–516.

Stephens, M., Smith, N. J. and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**, 978–989.

Tabor, H. K., Risch, N. J. and Myers, R. M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* **3**, 391–397.

Terwilliger, J. D. and Ott, J. (1992). A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum Hered* **42**, 337–346.

Thomas, D., Xie, R. and Gebregziabher, M. (2004). Two-stage sampling designs for gene association studies. *Genet Epidemiol* **27**, 401–414.

Thorleifsson, G., Walters, G. B., Gudbjartsson, D. F., Steinthorsdottir, V., Sulem, P., Helgadottir, A., Styrkarsdottir, U., Gretarsdottir, S., Thorlacius, S., Jonsdottir, I., Jonsdottir, T., Olafsdottir, E. J., Olafsdottir, G. H. et al. (2009). Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat Genet* **41**, 18–24.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B* **58**, 267–288.

Tzeng, J.-Y., Byerley, W., Devlin, B., Roeder, K. and Wasserman, L. (2003). Outlier detection and false discovery rates for whole-genome DNA matching. *Journal of the American Statistical Association* **98**, 236–246.

Tzeng, J.-Y., Devlin, B., Wasserman, L. and Roeder, K. (2003). On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am J Hum Genet* **72**, 891–902.

Van der Meulen, M. A. and te Meerman, G. J. (1997a). *Genetic Mapping of Disease*

Genes, chapter 9: Association and haplotype sharing due to identity by descent, with an application to genetic mapping, pages 115–135. Academic Press Ltd., London.

Van der Meulen, M. A. and te Meerman, G. J. (1997b). Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. *Genet Epidemiol* **14**, 915–920.

Wang, K., Li, M. and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* **81**.

Wang, T. and Elston, R. C. (2007). Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am J Hum Genet* **80**, 353–360.

Wang, W. Y. S., Barratt, B. J., Clayton, D. G. and Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* **6**, 109–118.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678.

Willer, C. J., Speliotes, E. K., Loos, R. J. F., Li, S., Lindgren, C. M., Heid, I. M., Berndt, S. I., Elliott, A. L., Jackson, A. U., Lamina, C., Lettre, G., Lim, N., Lyon, H. N., McCarroll, S. A., Papadakis, K., Qi, L., Randall, J. C. et al. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* **41**, 25–34.

Wrensch, M., Jenkins, R. B., Chang, J. S., Yeh, R.-F., Xiao, Y., Decker, P. A., Ballman, K. V., Berger, M., Buckner, J. C., Chang, S., Giannini, C., Halder, C. et al. (2009). Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. *Nat Genet* **41**, 905–908.

Wright, A. F., Carothers, A. D. and Pirastu, M. (1999). Population choice in mapping genes for complex diseases. *Nat Genet* **23**, 397–404.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E. and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**, 714–721.

Xiong, M., Zhao, J. and Boerwinkle, E. (2002). Generalized t^2 test for genome association studies. *Am J Hum Genet* **70**, 1257–1268.

Yu, K., Wang, Z., Li, Q., Wacholder, S., Hunter, D. J., Hoover, R. N., Chanock, S. and Thomas, G. (2008). Population substructure and control selection in genome-wide association studies. *PLoS One* **3**, e2551.

Zheng, G. and Tian, X. (2005). The impact of diagnostic error on testing genetic association in case-control studies. *Stat Med* **24**, 869–882.