Statistical Methods to Address the Challenges Posed by Rare Variants and Missing Genotypes in Case-Control Resequencing Studies

Daniel D. Kinnamon

University of Miami, ddkinnamon@gmail.com

Follow this and additional works at: https://scholarlyrepository.miami.edu/oa_dissertations

Recommended Citation
Kinnamon, Daniel D., "Statistical Methods to Address the Challenges Posed by Rare Variants and Missing Genotypes in Case-Control Resequencing Studies" (2013). Open Access Dissertations. 1004.
https://scholarlyrepository.miami.edu/oa_dissertations/1004
STATISTICAL METHODS TO ADDRESS THE CHALLENGES POSED BY RARE VARIANTS AND MISSING GENOTYPES IN CASE-CONTROL RESEQUENCING STUDIES

By

Daniel Davis Kinnamon

A DISSERTATION

Submitted to the Faculty of the University of Miami in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Coral Gables, Florida

May 2013
STATISTICAL METHODS TO ADDRESS THE CHALLENGESPOSED BY
RARE VARIANTS AND MISSING GENOTYPES IN
CASE-CONTROL RESEQUENCING STUDIES

Daniel Davis Kinnamon

Approved:

Eden R. Martin, Ph.D.
Professor of Human Genetics
M. Brian Blake, Ph.D.
Dean of the Graduate School

William K. Scott, Ph.D.
Professor of Human Genetics
Mitsunori Ogihara, Ph.D.
Professor of Computer Science

Miroslav Kubat, Ph.D.
Associate Professor of Electrical and Computer Engineering
Case-control resequencing studies are growing in popularity as investigators apply novel massively parallel sequencing technologies to existing case-control data sets. However, the sequence data generated by these studies present several daunting analytic challenges. The present study focuses on addressing the challenges posed by rare variants and missing genotypes when performing a test for association between a disease and a locus using data from a case-control resequencing study.

Association tests that pool minor alleles into a measure of burden at a locus have been proposed to address allelic heterogeneity in the presence of rare variants. However, such pooling tests are not robust to the inclusion of neutral and protective variants, which can mask the association signal from risk variants, and may not be robust to randomly missing genotypes. In contrast, methods for locus-wide inference using nonnegative single-variant test statistics are robust to both the inclusion of neutral and protective variants and randomly missing genotypes. Therefore, three existing methods for locus-wide inference using nonnegative single-variant test statistics were compared to two widely cited pooling tests under realistic conditions. Analytic results for a simple model with one rare risk and one rare neutral variant demonstrated that pooling tests are less powerful than even Bonferroni-corrected single-variant tests in most situations. These results were extended by Monte Carlo simulations using variants with realistic minor
allele frequency and linkage disequilibrium spectra, disease models with multiple rare risk variants and extensive neutral variation, and varying rates of randomly missing genotypes. In all scenarios considered, at least one existing method using nonnegative single-variant test statistics had power comparable to or greater than the two pooling tests considered. These results suggest that efficient locus-wide inference using single-variant test statistics should be reconsidered as a useful framework for addressing the challenge posed by rare variants in case-control resequencing studies.

Methods that perform efficient locus-wide inference using nonnegative single-variant test statistics also partially address the challenge posed by missing genotypes because they can use all available genotype data. When these methods are based on permutation tests, inferences will be valid if genotypes are randomly missing—that is, if the probability of a missing genotype at a variant does not depend on other observed or unobserved variables in the study. However, it was unclear whether methods based on permutation tests would yield valid inferences for nonrandomly missing genotypes. Therefore, a rigorous theoretical framework for constructing valid permutation tests was developed for genetic case-control studies with unrelated subjects and missing genotypes arising from a variety of missing data processes. The development began with the specification of a nonparametric probability model for the observed data in such a study. Group-theoretic arguments were then used to establish two conditions that together guarantee an exact level-α Monte Carlo permutation test for data generated under this nonparametric probability model. One of these conditions is not satisfied for the most frequently used Monte Carlo permutation test, and this test is guaranteed to be level α only for missing data processes with certain characteristics. An alternative Monte Carlo
permutation test, which is exact level $\alpha$ as long as all covariates influencing the missing data process are identified and recorded, was therefore proposed. The theoretical development was supplemented with Monte Carlo simulations for a variety of test statistics and missing data processes. These results demonstrate that Monte Carlo permutation tests must be constructed with careful consideration of the missing data process to adequately address the challenge posed by missing genotypes and avoid inferential errors.
To my wife, Mary, and my parents, David and Karen
ACKNOWLEDGEMENTS

I owe a debt of gratitude to my advisor, Dr. Eden R. Martin, for her years of excellent mentorship. She has provided me with sound guidance and ample opportunities for professional enrichment that have helped me to grow as both a scientist and a person. I also wish to thank my committee members—Drs. William K. Scott, Mitsunori Ogihara, and Miroslav Kubat—for their guidance and encouragement as I completed my research.

I am also grateful to Dr. Ray Hershberger for giving me the opportunity to collaborate with him early in my doctoral program. The candidate gene resequencing data that were part of that collaboration helped to inspire and anchor the statistical developments presented in this dissertation. My simulations would not have been possible without the computing infrastructure and support provided by the Center for Computational Science at the University of Miami and its High Performance Computing team led by Joel P. Zysman.

My scholarship and research were supported financially by the University of Miami Fellowship, the Dr. John T. Macdonald Foundation Department of Human Genetics at the University of Miami Miller School of Medicine, the GSA Merit Scholarship from the Graduate Student Association of the University of Miami, the Dr. Louis J. Elsas Research Award in Biochemical Genetics from the Dr. John T. Macdonald Foundation Department of Human Genetics, and Grant Number RC2HG005605 from the National Human Genome Research Institute.
# TABLE OF CONTENTS

LIST OF FIGURES ........................................................................................................... vi

LIST OF TABLES ............................................................................................................. vii

CHAPTER 1: INTRODUCTION ....................................................................................... 1
  1.1 Background ............................................................................................................. 1
  1.2 Motivating Example ............................................................................................ 2
  1.3 Addressing the Challenge Posed by Rare Variants ........................................... 4
  1.4 Addressing the Challenge Posed by Missing Genotypes .................................... 7

CHAPTER 2: EXISTING METHODS USING NONNEGATIVE SINGLE-VARIANT TEST STATISTICS AS ALTERNATIVES TO POOLING TESTS FOR SEQUENCE DATA WITH RARE VARIANTS ................................................................................... 10
  2.1 Summary ........................................................................................................... 10
  2.2 Characteristics of Variants in Actual Sequence Data ......................................... 10
  2.3 Impact of Neutral Variation in a Simple Model ............................................... 13
  2.4 Monte Carlo Simulations .................................................................................. 20
  2.5 Detailed Derivations ......................................................................................... 39

CHAPTER 3: VALID PERMUTATION TESTS FOR GENETIC CASE-CONTROL STUDIES WITH MISSING GENOTYPES ..................................................................... 52
  3.1 Summary ........................................................................................................... 52
  3.2 Nonparametric Probability Model ..................................................................... 52
  3.3 Group-Theoretic Construction of Exact Level-\(\alpha\) Monte Carlo Permutation Tests .................................................................................................................. 56
  3.4 Standard Permutation Test ............................................................................... 62
  3.5 Restricted Permutation Test ............................................................................. 65
  3.6 Monte Carlo Simulations .................................................................................. 67
  3.7 Detailed Derivations ......................................................................................... 79

CHAPTER 4: DISCUSSION ........................................................................................... 89
  4.1 Summary of Findings ....................................................................................... 89
  4.2 Addressing the Challenge Posed by Rare Variants ......................................... 91
  4.3 Addressing the Challenge Posed by Missing Genotypes ................................. 95

REFERENCES ............................................................................................................... 100
LIST OF FIGURES

Figure 2.1: MAF and within-gene pairwise LD distributions in actual sequence data .................................................................................................................. 12

Figure 2.2: Analytic power comparisons in a small sample .............................................................. 17

Figure 2.3: Analytic power comparisons in a moderate sample ...................................................... 18

Figure 2.4: Analytic power comparisons in a large sample ............................................................ 19

Figure 2.5: MAF and pairwise LD distributions in simulated sequence data ................................. 32

Figure 2.6: Simulated type I error rate comparison .................................................................... 33

Figure 2.7: Simulated power comparison for rare risk variants (MAF<0.005; OR=3) .................. 34

Figure 2.8: Simulated power comparison for rare risk variants (MAF<0.01; OR=2) .................... 35

Figure 2.9: Simulated power comparison for a mixture of rare and common risk variants ....... 36

Figure 3.1: Type I error rates of Monte Carlo permutation tests based on the CA max statistic ........................................................................................................ 74

Figure 3.2: Type I error rates of Monte Carlo permutation tests based on the CA sum statistic ........................................................................................................ 75

Figure 3.3: Type I error rates of Monte Carlo permutation tests based on the FW burden statistic ........................................................................................................ 76

Figure 3.4: CA max restricted permutation p-values versus standard permutation p-values ........................................................................................................ 78

Figure 3.5: CA sum restricted permutation p-values versus standard permutation p-values ........................................................................................................ 78

Figure 3.6: FW burden restricted permutation p-values versus standard permutation p-values ........................................................................................................ 79
LIST OF TABLES

Table 3.1: Summary of missing data processes used in the Monte Carlo simulations................................................................. 70

Table 3.2: Estimated power under a disease model with rare risk variants......................... 77
CHAPTER 1: INTRODUCTION

1.1 Background

The previous decade has witnessed a massive increase in the number of genome-wide association (GWA) studies. Since the first successful GWA study identified the association between a polymorphism in \textit{CFH} and age-related macular degeneration in 2005 [Klein et al., 2005], the number of published GWA studies has burgeoned to more than 1,500 [Hindorff et al., 2013]. As a result, large consortia spanning multiple institutions (e.g., the Wellcome Trust Case Control Consortium [Burton et al., 2007], the National Institute of Diabetes and Digestive and Kidney Diseases Inflammatory Bowel Disease Genetics Consortium [Barrett et al., 2008], the Alzheimer’s Disease Genetics Consortium [Naj et al., 2011], and the International Multiple Sclerosis Genetics Consortium [Sawcer et al., 2011]) have amassed large case-control data sets comprising hundreds or thousands of individuals.

Although GWA studies have identified many common susceptibility variants, these variants rarely explain more than 5-10\% of disease heritability [Manolio et al., 2009; Schork et al., 2009]. Moreover, a growing body of evidence suggests that many complex diseases, such as breast cancer, autism, and schizophrenia, follow the common disease/rare variant (CD/RV) model in which minor alleles at one or more rare variants in a locus influence disease risk [McClellan and King, 2010]. Because the majority of these minor alleles are unlikely to occur on the same haplotype as a particular allele of a

---

1 Portions of this chapter have been published previously as: Kinnaman DD, Hershberger RE, Martin ER. 2012. Reconsidering association testing methods using single-variant test statistics as alternatives to pooling tests for sequence data with rare variants. PLoS ONE 7:e30238.
common variant, it is unlikely that an association between rare variants and disease will appear in GWA study due to linkage disequilibrium (LD) alone [Bodmer and Bonilla, 2008; McClellan and King, 2010]. This fact, combined with the availability of novel massively parallel sequencing technologies, has led some authors to recommend interrogating this rare variation with whole-exome and whole-genome sequencing, particularly in the context of family-based studies [Manolio et al., 2009; McClellan and King, 2010]. As investigators look to leverage the large bodies of DNA samples and phenotype data previously collected for GWA studies, the number of studies applying whole-exome and whole-genome sequencing in case-control designs will likely grow rapidly. In fact, the first studies performing whole-exome sequencing on case-control data sets are already appearing in the literature (see, e.g., [Albrechtsen et al., 2013; Neale et al., 2012; Nuytemans et al., 2013]).

The present study considers the general problem of using data from a case-control resequencing study to perform a test for association between a disease and a locus, which is defined as a region of contiguous sequence including many variants (i.e., polymorphic sequence positions). These variants may be either common or rare; those with minor allele frequencies (MAFs) <2-3% [Bodmer and Bonilla, 2008] are termed rare variants, and those with higher MAFs are termed common variants. While this problem presents several daunting analytic challenges, the present study will focus on two specific challenges motivated by an actual case-control resequencing data set.

1.2 Motivating Example

A case-control resequencing study of six candidate genes for dilated cardiomyopathy (CSRP3, LDB3, MYH7, SCN5A, TCAP, and TNNT2) [Rampersaud et al.,
2010] provides an excellent example of the analytic challenges posed by such data. In this resequencing study, bi-directional Sanger sequencing of all coding sequence, at least 50 base pairs (bp) into 5'/3' untranscribed regions (UTRs), and at least 40 bp into all introns was used to scan a total of approximately 53 kilobases (kb) for variation. The data therefore represent the interrogation of a small region of mostly protein-coding sequence with an assay widely considered to be the gold standard.

The first major challenge was posed by the large number of rare variants. Of the 331 biallelic variants that were polymorphic in the case-control sample, 257 (78%) had a combined sample MAF\(\leq3\%\). Because each unique combination of single-variant alleles defines a unique haplotype, this implies substantial allelic heterogeneity and a large number of low-frequency haplotypes. Such allelic heterogeneity, in turn, may reduce the power of association tests [Martin, 2006; Pritchard, 2001; Schork et al., 2009].

The second major challenge was posed by the ubiquitous missing genotypes in these data. With a multivariate technique designed to address the challenge posed by allelic heterogeneity [Li and Leal, 2008], between 15% and 65% of individuals had to be excluded from analysis for each gene because only individuals with complete common variant genotype data at the gene could be used. Unfortunately, imputation was not a viable option for obtaining complete genotype data in every individual due to the paucity of sequence-level reference panels at the time of analysis. Because the excluded individuals had usable genotype data at some variants, the loss of power from the technique’s failure to address missing genotypes may have more than offset any power gains from its attempt to address allelic heterogeneity. Thus, this resequencing study demonstrates a clear need for case-control association tests designed to handle both
allelic heterogeneity in the presence of rare variants as well as missing genotypes, especially given that the challenges encountered are likely to be amplified by massively parallel sequencing of wider genomic regions.

1.3 Addressing the Challenge Posed by Rare Variants

To mitigate the power loss due to allelic heterogeneity in the presence of rare variants, several authors have suggested that pooling minor alleles into a measure of burden at a locus will be necessary [Li and Leal, 2008; Madsen and Browning, 2009; Schork et al., 2009]. Tests based on either collapsing rare variants in a locus into a single indicator of the presence of any minor alleles [Li and Leal, 2008; Morris and Zeggini, 2010] or summing weighted minor allele counts over rare, and sometimes also common, variants in a locus [Hoffmann et al., 2010; Madsen and Browning, 2009; Price et al., 2010; Sul et al., 2011] have been proposed as alternatives to single-variant tests. Techniques involving collapsing or summing will subsequently be referred to as pooling tests. Two of the earliest proposals, the Combined Multivariate and Collapsing (CMC) method [Li and Leal, 2008] for collapsing and the Weighted Sum Statistic (WSS) method [Madsen and Browning, 2009] for summing, are commonly used as benchmarks for novel methods in the rare variant association testing literature [Hoffmann et al., 2010; Ionita-Laza et al., 2011; Neale et al., 2011; Price et al., 2010; Sul et al., 2011; Wu et al., 2011] based on results suggesting that these techniques were superior to locus-wide inference using single-variant test statistics.

The power of pooling tests will depend on the linkage disequilibrium (LD) patterns in sequence data. Simulations using a coalescent approximation to a neutral two-locus Wright-Fisher infinite allele model have shown that a substantial proportion of the
pairwise LD between biallelic variants can be expected to be negative (i.e., $D<0$), even at very high levels of recombination [Hudson, 1985]. Most importantly for sequence data, negative pairwise LD values become more likely when including variants with relatively rare minor alleles [Hudson, 1985]. To the extent that there is negative LD between neutral and risk variants within a locus, higher MAFs at a small number of risk variants in cases will be accompanied by higher MAFs at a larger number of neutral variants in controls. In this situation, case-control differences in the MAFs at individual risk variants may actually be masked by those at a large number of neutral variants in locus-wide summaries based on collapsing or summing minor alleles over all variants at a locus. Masking, in turn, reduces the power of pooling tests by obscuring locus-wide case-control differences. Such masking and power loss will be exacerbated by the inclusion of protective variants.

Because pooling tests lose power when neutral and protective variants are included, one sensible approach is to try to exclude such variants a priori by filtering on annotation and functional predictions. However, making such exclusions with high sensitivity and specificity will be difficult, particularly in non-coding regions for which little information is available. Even in coding regions, functional predictions may lead researchers astray. For example, recent studies have implicated synonymous variants in altering the function of protein products [Kimchi-Sarfaty et al., 2007] and causing disease [Brest et al., 2011]. Thus, methods for locus-wide inference that are inherently robust to the inclusion of neutral and protective variants are desirable.

Nonetheless, many novel methods have not departed from the pooling test framework but rather attempted to devise improved weighting schemes and adaptive
thresholds that reduce the influence of neutral and protective variants [Hoffmann et al., 2010; Price et al., 2010; Sul et al., 2011; Wu et al., 2011]. Despite deriving their approaches within fundamentally different frameworks, some of the newest methods that are robust to the inclusion of neutral and protective variants have arrived at procedures equivalent to performing locus-wide inference using nonnegative single-variant test statistics. For example, the Sequence Kernel Association Test (SKAT) [Wu et al., 2011] and the C-alpha test [Neale et al., 2011] base inference on weighted and unweighted sums of squared single-variant score statistics, respectively [Wu et al., 2011]. Under an additive genetic model without covariates and with variant weights equal to the inverse of the estimated null variance of each single-variant score statistic, the SKAT statistic is simply the sum of single-variant Cochran-Armitage trend chi-square statistics. Sums of nonnegative single-variant test statistics have previously been recommended as powerful methods for joint inference over multiple variants in candidate gene association studies [Chapman and Whittaker, 2008] and GWA studies [Hoh and Ott, 2003; Hoh et al., 2001]. The equivalence of new developments to existing methods for efficient locus-wide inference using nonnegative single-variant test statistics argues for broader investigation of other existing methods falling within this framework, such as permutation inference on the maximum single-variant Cochran-Armitage trend chi-square statistic in the locus.

In general, existing methods for locus-wide inference using nonnegative single-variant test statistics are inherently robust to the inclusion of neutral and protective variants. This robustness arises from the fact that the locus-wide inference depends on only the magnitude of the deviation from the null hypothesis at each variant and not the direction. Joint inference can be performed efficiently by using permutation, which, by
simulating draws from the joint null distribution of the single-variant test statistics, avoids conservative approximations (e.g., the Bonferroni correction) and accounts for LD-induced correlations between test statistics [Westfall and Young, 1993]. Existing methods that perform efficient locus-wide inference on nonnegative single-variant test statistics using permutation can therefore combine information across variants in a locus without masking. Such methods may also be able to extract additional association signal from neutral variants in negative LD with risk variants as well as protective variants. For these reasons, the first aim of the present study is to determine whether three existing methods for locus-wide inference using nonnegative single-variant test statistics are more powerful than two widely cited pooling tests in realistic sequence data with substantial neutral variation and randomly missing genotypes.

1.4 Addressing the Challenge Posed by Missing Genotypes

Methods that perform efficient locus-wide inference using single-variant test statistics also partially address the challenge posed by missing genotypes because they can use all available genotype data, which is not necessarily true of pooling tests. When these methods are based on permutation tests, inferences will be valid if genotypes are randomly missing—that is, if the probability of a missing genotype at a variant does not depend on other observed or unobserved variables in the study. However, it is unclear whether methods based on permutation tests will yield valid inferences for nonrandomly missing genotypes.

To illustrate the problem, consider a hypothetical genetic case-control study in which affection status is always observed but genotypes at some variants are missing. The usual permutation test for genetic case-control studies with unrelated subjects
considers all possible permutations of affection status indicators, or equivalently intact observed genotype vectors, among subjects. The test statistic is recalculated for each permutation, and the null hypothesis of no association is then rejected if the p-value calculated using the discrete uniform distribution over all permutations is less than or equal to a nominal $\alpha$. If the genotypes were randomly missing for all subjects, then every observed genotype vector would be an independent realization from the same distribution. In that case, every permutation of the observed genotype vectors would have the same probability, which would mean that the standard permutation test would have a type I error rate of at most $\alpha$ [Good, 2000; Pesarin and Salmaso, 2010]. However, if genotypes were missing at different rates across individuals because of differing data collection or assay techniques, the observed genotype vectors would come from different distributions despite the fact that the true genotypes were identically distributed. Therefore, all permutations of the observed genotype vectors might not have the same probability, and the usual permutation test might have a type I error rate exceeding $\alpha$.

The primary attraction of permutation tests is that, if correctly constructed for the testing problem at hand, they are exact level $\alpha$; that is, the type I error rate is guaranteed to be at most the nominal $\alpha$ for all test statistics and underlying data generating processes consistent with the null hypothesis, regardless of the sample size [Good, 2000; Lehmann and Romano, 2005; Pesarin and Salmaso, 2010]. Nonetheless, it is difficult to find any thorough exploration of how to construct exact level-$\alpha$ permutation tests in the presence of missing genotypes. Some authors have considered permutation tests with missing data [Naddeo, 2002; Pesarin and Salmaso, 2010], but their formulations of the testing problem require assuming that, conditional on being observed, genotypes are identically
distributed in cases and controls for all missing data processes under a null disease model. Other authors provide solutions that rely on modified test statistics or data transformations to address missing data in specific applications [Entsuah, 1990; Good, 2000]. Previous studies in other fields have also presented empirical finite-sample results regarding the performance of specific types of permutation tests with missing values [Kennes et al., 2012; Mundry, 1999]. These developments, while useful, are difficult to generalize to the plethora of possible test statistics and data generating processes for a case-control resequencing study.

The question of how to construct exact level-\(\alpha\) permutation tests in the presence of missing genotypes is extremely important for the analysis of case-control resequencing studies. In fact, many pooling tests and other novel methods to address allelic heterogeneity in sequence data also rely on permutation tests (see, e.g., [Hoffmann et al., 2010; Ionita-Laza et al., 2011; Madsen and Browning, 2009; Neale et al., 2011; Price et al., 2010]). Therefore, the second aim of the present study is to develop a broadly applicable method for constructing exact level-\(\alpha\) permutation tests for genetic case-control studies in the presence of missing genotypes.
2.1 Summary

In this chapter, three existing methods for locus-wide inference using nonnegative single-variant test statistics are compared to the two originally proposed pooling tests, the CMC [Li and Leal, 2008] and WSS [Madsen and Browning, 2009] tests, under realistic conditions. Analytic power calculations under a simple model are used to demonstrate that pooling tests may have lower power than even Bonferroni-corrected single-variant tests in the presence of neutral variants with realistic patterns of LD. Simulations based on variants with realistic MAF and LD spectra are then used to extend the basic conclusions of this simple model to more complex situations with allelic heterogeneity, extensive neutral variation, and randomly missing genotype data.

2.2 Characteristics of Variants in Actual Sequence Data

To ground the interpretation of analytic power approximations in actual data and provide a basis for evaluating the realism of simulated data, MAF and LD distributions were estimated for variants in resequencing data from six genes (CSRP3, LDB3, MYH7, SCN5A, TCAP, and TNNT2). These data were obtained from a previous study of candidate genes for dilated cardiomyopathy [Rampersaud et al., 2010]. The six genes spanned a total of 236,059 bp, of which 53,466 bp were scanned for variation. These data should provide a useful snapshot of sequence-level variation within protein-coding genes.

---

2 Portions of this chapter have been published previously as: Kinnamon DD, Hershberger RE, Martin ER. 2012. Reconsidering association testing methods using single-variant test statistics as alternatives to pooling tests for sequence data with rare variants. PLoS ONE 7:e30238.
of a wide range of sizes (3 kb – 103 kb), numbers of exons (2 – 40), and chromosomal locations (1q32, 3p21, 10q22.3-23.2, 11p15.1, 14q12, and 17q12). Bi-directional Sanger sequencing of all coding sequence, at least 50 bp into 5’-/3’-UTRs, and at least 40 bp into all introns was performed by SeattleSNPs under contract to the National Heart, Lung, and Blood Institute resequencing service. The data set comprised 184 unrelated controls of European descent from the Coriell database with variant call rates ≥80% after removal of low-quality variants called in <80% of individuals in the study.

A total of 215 biallelic variants, including both single-nucleotide polymorphisms (SNPs) and small insertions/deletions, were identified, yielding approximately 4 variants per kb scanned. There was no evidence against Hardy-Weinberg equilibrium (HWE) (Monte Carlo exact \( P \geq 0.001 \)) at 211 of these variants, which were carried forward to the analysis of MAF and LD distributions. Pairwise LD measured by the correlation coefficient between major alleles, \( r \), was calculated only between variants within the same gene under the assumption of HWE using the method of Weir and Cockerham [Weir and Cockerham, 1979] as implemented in PROC ALLELE, SAS/GENETICS, Version 9.2 (SAS Institute Inc., Cary, NC). This correlation coefficient is the same as the correlation coefficient between minor alleles for a biallelic variant.

More than half of variants had MAFs below 0.01, confirming that a multitude of rare variants is likely to be a distinguishing characteristic of sequence data (Figure 2.1, Panel A). In addition, the majority of pairwise LD between variants within the same gene was small and negative, with more than 75% of \( r \) values below 0 (Figure 2.1, Panel B). Pairwise LD between rare variants with MAF ≤0.01 was even more concentrated in small
negative values, with 95% of values falling between \( r = -0.0095 \) and \( r = -0.0028 \). These results confirmed in actual data predictions regarding the sampling distribution of LD based on coalescent theory [Hudson, 1985].

Although the negative pairwise \( r \) values between variants with MAF≤0.01 may seem small in magnitude, they are not inconsistent with negative LD having a substantial impact on pooling tests. First, the theoretical minimum for \( r \) between variants with MAF≤0.01 is \( r_{\text{min}} = -0.0101 \), so many of these \( r \) values may actually represent \( D' \) values near 1 that would be considered strong LD. Second, because neutral variants are far more numerous than risk variants in the genome, an appreciably higher MAF at a single risk variant in cases can mean slightly higher MAFs at numerous neutral variants in controls when most \( r \) values are negative. If truly neutral variants are not detected with high sensitivity and filtered out prior to analysis, the cumulative case-control MAF difference over this large number of neutral variants can easily mask the cumulative case-control
MAF difference over a few risk variants. Therefore, small negative pairwise $r$ values between rare variants can have an appreciable effect on pooling tests.

### 2.3 Impact of Neutral Variation in a Simple Model

The Cochran-Armitage test for a single variant is well-known in the field of genetics [Freidlin et al., 2002; Sasieni, 1997; Slager and Schaid, 2001]. It possesses some desirable properties, including robustness to departures from HWE [Sasieni, 1997] and ease of calculation, that make it widely applicable. Suppose that $R$ cases and $S$ controls are independently sampled, and let $N = R + S$. It will be convenient for subsequent exposition to present the test statistic $T$ as a generalization of the form presented in Freidlin et al. [Freidlin et al., 2002] to multi-variant genotypes:

$$U = \sum_k X_k \left( \frac{S}{N} r_k - \frac{R}{N} s_k \right)$$

$$\text{Var}_{H_0}(U) = N\hat{\sigma}^2_0 = \frac{RS}{N^3} \left( N \sum_k X^2_k n_k - \left( \sum_k X_k n_k \right)^2 \right)$$  

$$T = U^2 / \text{Var}_{H_0}(U) = U^2 / N\hat{\sigma}^2_0$$  

In (2-1), $X_k$ is a score corresponding to the $k^{th}$ single- or multi-variant genotype, denoted $G_k = [G_{1k}, \ldots, G_{vk}]$, comprising $v$ biallelic variants indexed by $j$. Also, let $r_k$ be the number of cases with genotype $G_k$, $s_k$ be the number of controls with genotype $G_k$, and $n_k = r_k + s_k$. The single-variant genotype at biallelic variant $j$ in multi-variant genotype $G_k$ is coded as $G_{jk}=0$, 1, or 2 minor alleles. Thus, multi-variant genotype $k$ denotes one unique combination of single-variant genotypes (e.g., $G_k=[0,0,1,2]$ for $v=4$ and one particular $k$) in the set of all $w=3^v$ possible combinations at each of the $v$ variants (i.e., $k=1,2,\ldots, 3^v-1,3^v$). With a single variant and scores equaling the number of minor alleles at that
variant, it is immediate that $G_k = [G_{1k}]$, $k=1, 2, \text{ or } 3$, and $X_k = G_{1k}$, so $T$ in (2-1) reduces to the single-variant Cochran-Armitage trend chi-square statistic.

Consider a simple model of a locus comprising one rare neutral ($j=1$) and one rare risk ($j=2$) biallelic variant with the same MAF ($p$). In this model, frequencies of haplotypes $[h_1, h_2]$ for a given level of LD ($D$) are calculated as $(1-p)^2+D$ for haplotype [0,0], $(1-p)p-D$ for haplotypes [0,1] and [1,0], and $p^2+D$ for haplotype [1,1], where 1 denotes the minor allele and 0 the major allele at either variant. Population frequencies of multi-variant genotypes $G_k = [G_{1k}, G_{2k}]$, denoted by $p_k$, are then determined under HWE. Disease risk is given by a multiplicative model, $P(A|G_k) = f_0 e^{\gamma z}$, with $A$ or $C$ denoting affection status (affected case or control), $f_0$ denoting the penetrance for the variant 2 major allele homozygote, and $\gamma$ denoting the relative risk for an additional minor allele at variant 2. The conditional frequencies of each $G_k$ in cases ($p_{kA}$) and controls ($p_{kC}$) are then determined based on this risk model and population multi-variant genotype frequencies using Bayes’ rule:

$$p_{kA} = P(A|G_k) p_k / \sum_k P(A|G_k) p_k$$
$$p_{kC} = [1 - P(A|G_k)] p_k / \sum_k [1 - P(A|G_k)] p_k$$

Under the generalization in (2-1) and the assumed sampling model, $(r_1, \ldots, r_w) \sim \text{Multinomial}(R; p_{1A}, \ldots, p_{wA})$ and $(s_1, \ldots, s_w) \sim \text{Multinomial}(S; p_{1C}, \ldots, p_{wC})$.

The null hypothesis that no variant in the locus influences disease risk implies that $P(A|G_k) = f_0$ and $p_{kA} = p_{kC} = p_k$ for all $k$. Under the alternative hypothesis, at least one variant in the locus influences disease risk, implying that $p_{kA} \neq p_{kC}$ for some $k$. 
Freidlin et al. [Freidlin et al., 2002] provide formulas for the variance under this null hypothesis as well as the expectation and variance under this alternative hypothesis for the statistic $U$ in (2-1) with arbitrary scores $X_k$ when $v=1$. These formulas immediately generalize to multinomial genotype distributions with more than 3 possible genotypes ($w>3$):

$$E_{H_0}[U] = N\mu_0 = 0$$
$$\text{Var}_{H_0}(U) = N\sigma_0^2 = N\left\{\frac{RS}{N^2}\left[\sum_k X_k^2 p_k - \left(\sum_k X_k p_k\right)^2\right]\right\}$$
$$E_{H_a}[U] = N\mu_a = N\left\{\frac{RS}{N^2}\sum_k X_k \left(p_{k|a} - p_{k|C}\right)\right\}$$
$$\text{Var}_{H_a}(U) = N\sigma_a^2 = N\left\{\frac{RS^2}{N^3}\left[\sum_k X_k^2 p_{k|a} - \left(\sum_k X_k p_{k|a}\right)^2\right]\right\} + \frac{R^2S}{N^3}\left[\sum_k X_k^2 p_{k|C} - \left(\sum_k X_k p_{k|C}\right)^2\right]\right\} + \mu_a^2$$

Under the null hypothesis, the asymptotic distribution of $U$ is $N(0,N\sigma_0^2)$, and $\hat{\sigma}_0^2$ defined in (2-1) converges in probability to $\sigma_0^2$ [Freidlin et al., 2002]. Thus, $T$ has the same asymptotic $\chi^2$ null distribution as $U^2/N\sigma_0^2$ by Slutsky’s Theorem because $\sigma_0^2/\hat{\sigma}_0^2$ converges in probability to 1 [Casella and Berger, 2002]. Under the alternative hypothesis, the asymptotic distribution of $U$ is $N(N\mu_a,N\sigma_a^2)$ and $\hat{\sigma}_0^2$ converges in probability to [Freidlin et al., 2002]:

$$\hat{\sigma}_a^2 + \mu_a^2 = \frac{R^2S}{N^3}\left[\sum_k X_k^2 p_{k|a} - \left(\sum_k X_k p_{k|a}\right)^2\right] + \frac{RS^2}{N^3}\left[\sum_k X_k^2 p_{k|C} - \left(\sum_k X_k p_{k|C}\right)^2\right] + \mu_a^2$$

Thus, the asymptotic power function for any two-sided test based on $T$ at nominal level $\alpha$ is:
\[ \beta(\alpha, N, \mu, \sigma^2, \sigma_a^2) = P_{H_0}(T > \chi^2_{1, \alpha}) = P_{H_0}\left( \frac{U^2}{N\sigma_0^2} > \chi^2_{1, \alpha} \right) = P_{H_0}\left( \frac{U^2}{N\sigma_0^2} \frac{\sigma_a^2}{\sigma_0^2} > \chi^2_{1, \alpha} \right) \]

\[ \approx P_{H_0}\left( \frac{U^2}{N\sigma_0^2} \frac{\sigma_a^2 + \mu_a^2}{\sigma_a^2} > \chi^2_{1, \alpha} \right) = P_{H_0}\left( \frac{U^2}{N\sigma_0^2} \frac{\sigma_a^2 + \mu_a^2}{\sigma_a^2} \chi^2_{1, \alpha} \right) \]

\[ = 1 - F_{\chi^2}(N\mu_a^2/\sigma_a^2) \left( \frac{\sigma_a^2 + \mu_a^2}{\sigma_a^2} \chi^2_{1, \alpha} \right) \]

where \( F_{\chi^2}(N\mu_a^2/\sigma_a^2) \) refers to the noncentral \( \chi^2 \) CDF with 1 degree of freedom and noncentrality parameter \( N\mu_a^2/\sigma_a^2 \) [Timm, 2002].

Particular choices of scores \( X_k \) in the statistic \( T \) yield a single-variant Cochran-Armitage trend test, a collapsing test, and a summing test (see Subsection 2.5.1). This correspondence allows the use of the generalized formulas presented above to approximate the locus-wide asymptotic power of each approach to association testing under the simple model described earlier. The single-variant trend test uses \( X_k = G_{jk} \) for variant \( j \); collapsing uses \( X_k = I\left[ \sum_j G_{jk} = 0 \right] \), where \( I[\text{expression}] = 1 \) if the expression in brackets is true and 0 otherwise; and summing uses \( X_k = \sum_j G_{jk} \). A lower bound for the locus-wide power function of Bonferroni-corrected inference on the maximum single-variant Cochran-Armitage trend chi-square statistic in the locus (the BC-CA test) is used because calculating the joint distribution of the single-variant test statistics in the presence of LD is analytically intractable. This lower bound is simply the power of the Bonferroni-corrected Cochran-Armitage trend test for the risk variant alone (see Subsection 2.5.2).
Figures 2.2-2.4 plot power as a function of $r$ for $f_0 = 0.05$ and two types of rare variant pairs ($p = 0.005$ and $\gamma = 3$; $p = 0.01$ and $\gamma = 2$). Balanced case-control samples with total sizes of $N = 500$, 1,000, and 2,000 were considered. Values of $r$ were chosen by taking 100 evenly spaced increments of $D$ starting from $D_{\min} = -p^2$ and ending at $D_{\max} = p(1-p)$.

Although the power of the collapsing and summing tests exceeded the lower bound for the BC-CA test with larger positive $r$ values in small samples ($N=500$), the worst-case power of the BC-CA test was greater than that of the collapsing and summing tests for all $r \leq 0.08$ under both models considered (Figure 2.2). Moreover, the BC-CA test had a power advantage over an even larger range of $r$ values under the same models in moderate ($r \leq 0.34$) and large ($r \leq 0.52$) samples (Figures 2.3 and 2.4). In the actual
sequence data in Section 2.2, over 95% of r values between variants with MAFs in the range considered in these figures (≤0.01) fell below 0, suggesting that the BC-CA test should enjoy a power advantage in most practical situations.

The relationship between the power of the three tests and r can be better understood by first considering the properties of the BC-CA, collapsing, and summing tests when r=1. In this situation, the number of minor alleles at the neutral variant must always equal the number of minor alleles at the risk variant because minor alleles at both variants must appear on the same haplotype. Thus, the multi-variant genotype frequencies are the same as the genotype frequencies at the risk variant alone, and the scores are $X_k=1, 0, \text{ or } 0$ (collapsing) or $X_k=0, 2, \text{ or } 4$ (summing) when $G_{2k}=0, 1, \text{ or } 2$, respectively. Under these circumstances, the collapsing and summing tests are equivalent to level-$\alpha$ single-variant Cochran-Armitage tests for the risk variant using scores for a dominant
A model and an additive model, respectively. Because these tests are not Bonferroni corrected, their power when $r=1$ is substantially above the lower bound for the BC-CA test, which is based on a level-$\alpha/2$ single-variant Cochran-Armitage test for the risk variant using scores for an additive model.

However, as $r$ decreases, the amount of noise introduced into the collapsing and summing test statistics by including the neutral variant increases and results in a concomitant decrease in power. Moreover, the problem of masking by the rare neutral variant further reduces power when $r<0$. The worst-case power of the BC-CA test, which inefficiently combines single-variant test statistics, was substantially greater than the power of the collapsing and summing tests for $r<0$, which is where over 95% of the $r$ values between variants with MAF≤0.01 fell in the actual sequence data in Section 2.2. These results suggest that, by eliminating the problems of noise and masking, even inefficient techniques for locus-wide inference using nonnegative single-variant test

---

**Figure 2.4: Analytic power comparisons in a large sample**

Analytic locus-wide power at $\alpha=0.05$ of the BC-CA (lower bound), collapsing, and summing tests at a locus comprising one neutral and one risk variant as a function of the pairwise correlation coefficient between major/minor alleles ($r$). The variants had the same MAF=0.005 (Panel A) or MAF=0.01 (Panel B), and the relative risk was 3 (Panel A) or 2 (Panel B) for each additional minor allele at the risk variant. Both panels assume penetrance of 0.05 for the major allele homozygote at the risk variant and a balanced case-control sample with $N=2,000$ total subjects.
statistics can yield more powerful tests for association than pooling minor alleles in the presence of rare neutral variants.

2.4 Monte Carlo Simulations

Monte Carlo simulations were used to extend the conclusions drawn from the two-variant locus model to a larger locus with heterogeneous risk alleles, extensive neutral variation, realistic LD patterns, and randomly missing genotype data. The major components of the simulation procedure will be outlined briefly here. One thousand populations of haplotypes at a 100 kb locus with realistic MAF and LD spectra were first generated based on a neutral coalescent model. Using the same 1,000 haplotype populations, a separate simulation was then conducted for each combination of user-specified risk variant parameters and sample size as follows:

1) The disease model for each haplotype population was generated by randomly selecting risk variants for inclusion based on user-specified parameters.

2) A case-control sample was drawn from each haplotype population according to the disease model and a user-specified sample size.

3) Data sets with randomly missing genotypes were generated from each sample for each user-specified call rate.

4) All association testing techniques were applied to each data set.

Type I error and power for each technique were estimated for balanced case-control samples of total sizes $N=500, 1,000,$ and $2,000$. The disease risk for the multi-variant genotype with no minor alleles at any risk variant was 5% for all simulations. For power, 50 risk variants with independent effects were randomly selected in each haplotype population under three different disease models:
1) Multiple rare risk variants (MAF<0.005; odds ratio (OR)=3)

2) Multiple rare risk variants (MAF<0.01; OR=2)

3) Combinations of multiple rare risk variants (MAF<0.01; OR=2), low-frequency risk variants (0.01≤MAF<0.05; OR=1.5), and common risk variants (0.05≤MAF<0.10; OR=1.2)

The first two models represent situations in which pooling tests are expected to perform best, and the third model is included to consider cases where both common and rare variants might contribute to disease susceptibility. The chosen number of risk variants represented ~5% of all variants at the locus in the average haplotype population, which reflects a situation in which associations between the locus and disease were driven by heterogeneous risk alleles characterized by a small number of risk variants among a much larger number of neutral variants. Per-position genotype call rates of 100% (complete data), 99.5%, and 95% were simulated. In the following subsections, a detailed description of each component of the simulation procedure is provided.

2.4.1 Haplotype Populations

One thousand populations of 10,000 haplotypes each were generated at a 100 kb locus, which is representative of a larger human protein coding gene based on recent data (mean size: 27 kb, range: 1 kb – 2,400 kb) [Antonarkis, 2010]. Haplotype populations were generated according to a standard neutral coalescent approximation to the Wright-Fisher model with a finite-sites recombination model and an infinite-sites mutation model, which is most accurate when the number of haplotypes sampled is small relative to the number of haplotypes in the population and the recombination rate between adjacent bases is small [Hudson, 1983, 2002]. The program MS [Hudson, 2002] was used
with parameters corresponding to a per-nucleotide neutral mutation rate of $2.5 \times 10^{-8}$ estimated assuming an effective diploid population size of $N_e \approx 10,000$ [Nachman and Crowell, 2000], a recombination rate of $1 \times 10^{-8}$ between adjacent nucleotides derived by using the approximation $1 \text{ cM} \approx 1 \text{ Mb}$ [Thomas, 2004], and $N_e = 10,000$ (i.e., 20,000 haplotypes). Following the suggestion in the MS documentation, the standard C random number generator was replaced with the well-known and highly robust Mersenne-Twister random number generator [Matsumoto and Nishimura, 1998]. All variants generated by MS are biallelic because it assumes the infinite-sites model of mutation.

2.4.2 Disease Model

Disease risk for the multi-variant genotype $G_k$ was determined according to a logistic model of the form:

$$P(A|G_k) = \left(1 + \exp\left(-\alpha - G_k \mathbf{B}\right)\right)^{-1}$$  \hspace{1cm} (2-2)

where $\alpha = \ln\left(f_o/(1 - f_o)\right)$, the log odds of the disease risk for the multi-variant genotype with no minor alleles at any risk variant, and $\mathbf{B} = [\beta_1, ..., \beta_l]^T$ is the vector of log odds ratios for the variants in the haplotype population. The odds ratio $\theta_j = \exp(\beta_j)$ reflects the increase in the odds of disease for each additional minor allele at variant $j$. This model implicitly assumes that (1) each additional minor allele has a multiplicative effect on the odds of disease and (2) this effect at variant position $j$ is independent of the effects at other variant positions.

To parameterize this model, the desired number of risk variants as well as a set of risk variant classes indexed by $c$ were specified. Class $c$ was defined by a half-open MAF
range, \([p_c^l, p_c^u]\), and an associated odds ratio, \(\theta_c\). For each haplotype population, the vector of log odds ratios \(B\) was populated by repeating the following steps until the specified number of risk variants was selected:

1) A variant \(j\) was randomly selected from among the variants in the haplotype population.
2) If the randomly selected variant \(j\) had a population MAF in the interval \([p_c^l, p_c^u]\) specified for risk variant class \(c\) and had not already been designated a risk variant, then it was labeled a risk variant and assigned a coefficient

\[
\beta_j = \ln(\theta_c).
\]

This procedure effectively randomly samples risk variant classes from the haplotype population in proportion to the occurrence of each MAF range in the population. All neutral variants had \(\beta_j = 0\) (\(\theta_j = 1\)).

2.4.3 Case-Control Samples

After the vector of log odds ratios \(B\) was populated for a haplotype population, a case-control sample was generated according to the disease model. To generate a case-control sample, the following procedure was repeated until the user-specified numbers of case and control subjects were selected:

1) Haplotypes were randomly drawn with replacement to form an individual’s multi-variant genotype \(G_k\).
2) The disease risk of the individual’s multi-variant genotype, \(P(A \mid G_k)\), was then calculated according to the logistic model in (2-2).
3) The individual was randomly assigned affection status \( A \) with probability 
\[
P(A|G_k)\] or \( C \) with probability \( 1-P(A|G_k) \). 

This method of forming \( G_k \) by randomly drawing haplotypes with replacement implicitly assumes random mating in the population.

2.4.4 Missing Genotypes

For each case-control sample, data sets with different rates of randomly missing genotypes were generated based on user-specified per-base-pair call rates. The observation process over the sequence was modeled as a two-state Markov chain with states “observed” (\( O \)) or “missing” (\( M \)) at each position defined by a single base pair. Given a call rate of \( \lambda < 1 \) per base pair, the number of base pairs that a chain remains in \( M \) before a genotype is called is distributed \( \text{Exp}(\lambda) \), assuming a sufficiently long sequence for continuous measurement of base pair position to be reasonable. At the position of this called genotype, the state of the chain changes from \( M \) to \( O \) with probability 1. Because genotypes are missing at a rate of \( 1-\lambda \) per base pair, the number of base pairs that the chain remains in \( O \) before a genotype is missing is distributed \( \text{Exp}(1-\lambda) \), and the chain transitions from \( O \) to \( M \) with probability 1 at this position. If position is rescaled to \([0, 1]\) by measuring in units of \( L \) base pairs, where \( L \) is the total sequence length, the distance the chain remains in \( M \) is distributed \( \text{Exp}(L\lambda) \), and the distance the chain remains in \( O \) is distributed \( \text{Exp}(L(1-\lambda)) \). It can be shown that the expected proportion of the sequence length that a Markov chain with these transition rates and probabilities spends in \( O \) is simply the call rate, \( \lambda \). Thus, for each call rate, the following steps were performed in each individual to generate the observed genotype data:
1) Starting at position 0 in $O$, a series of alternating $O$ and $M$ intervals on the [0,1] scale were generated according to the exponential transition distance distributions for the two-state Markov chain with call rate $\lambda$ over a sequence of length $L$.

2) The genotypes of variants with [0,1] scaled sequence positions not falling within an $O$ interval were set to missing.

The observation process defined by this Markov chain is independent of both affection status and the underlying genotypes, meaning that genotypes are randomly missing.

### 2.4.5 Association Testing

For each data set, several methods for association testing were applied. To ensure comparability across data sets from a population, the minor allele was determined based on the allele frequencies in the haplotype population from which the case-control sample was drawn. Variants that were monomorphic in a given data set were excluded from the test. A test based on a $p$-value rejected the null hypothesis in a data set at level $\alpha$ if the $p$-value was less than or equal to $\alpha$.

Single-variant Cochran-Armitage trend chi-square statistics were calculated as $T = N^O \rho^2$, where $N^O$ is the number of individuals with observed genotypes at the variant and $\rho$ is the Pearson correlation coefficient between the number of minor alleles at the variant and an affection status indicator equaling 1 for cases and 0 for controls across individuals with observed genotypes. Locus-wide inference was then performed using these single-variant statistics in accordance with three established methods.

The most widely known method is the BC-CA test presented above, which uses a conservative approximation that does not make efficient use of the single-variant
information for joint inference [Westfall and Young, 1993]. This test was implemented by rejecting the null hypothesis in a replicate at level $\alpha$ if the maximum Cochran-Armitage trend chi-square statistic (max $T$) in the locus was greater than or equal to the Bonferroni-corrected quantile of the asymptotic null $\chi^2$ distribution. This Bonferroni-corrected quantile was determined separately for each data set as $\chi^2_{1,1-\alpha/v}$, where $v$ is the number of polymorphic variants in the data set.

A second popular method involves performing locus-wide inference based on the permutation null distribution of max $T$ [Chapman and Whittaker, 2008], which is efficient because it does not use a conservative approximation and accounts for the LD-induced correlations between the single-variant $T$ values [Westfall and Young, 1993]. It has also demonstrated consistently good performance relative to other locus-wide tests in simulations of candidate gene SNPs with realistic LD [Chapman and Whittaker, 2008].

The permutation null distribution of max $T$ was obtained by repeatedly randomly shuffling affection status indicators, calculating all single-variant $T$ values, and recording the resulting value of the max $T$ statistic. Letting $Q_t$ denote the value of the max $T$ statistic in permutation $t$ and $Q_{obs}$ denote the observed value in the sample, the two-sided p-value was calculated from $m$ permutations as [Davison and Hinkley, 1997]:

$$\hat{P}_Q = \sum_{t=1}^{m} I[Q_t \geq Q_{obs}] + 1 \over m + 1$$  \hspace{1cm} (2-3)

Assuming that missingness does not depend on the underlying genotype or affection status, the Monte Carlo procedure described above will correctly estimate the permutation null distribution and yield a valid p-value. A Monte Carlo estimate of the
two-sided p-value was obtained from (2-3) with \( m = 10,000 \) permutations. The test based on the permutation p-value of the max \( T \) statistic will be referred to as the CA max test.

A final method involves performing locus-wide inference based on the permutation null distribution of the sum of Cochran-Armitage trend chi-square statistics (sum \( T \)) over the locus, which is also efficient because LD-induced correlations between single-variant \( T \) values are fully taken into account in the permutation null distribution. Variations on this theme have been proposed for candidate gene association studies [Chapman and Whittaker, 2008] and GWA studies [Hoh and Ott, 2003; Hoh et al., 2001]. Simulations of candidate gene SNPs with realistic LD found that the approach based on Fisher’s method for combining p-values, which is equivalent to a sum of nonnegative single-variant test statistics, performed well relative to other multi-SNP approaches when there were many variants in high LD [Chapman and Whittaker, 2008]. A Monte Carlo estimate of the two-sided p-value for the sum \( T \) statistic was obtained using the same permutation procedure as for the CA max test. The test based on the permutation p-value of the sum \( T \) statistic will be referred to as the CA sum test.

The CA sum test is also closely related to the SKAT and C-alpha test. Let \( U_j \) be the score statistic \( U \) in (2-1) for a single variant \( j \) with additive scores \( X_k = G_{jk} \), and let \( \omega_j \) be a pre-specified weight for variant \( j \). In the absence of covariates and with complete genotype data, the SKAT statistic can be expressed as \( Q_{SKAT} = \sum_{j=1}^{v} \omega_j U_j^2 \) (see Subsection 2.5.3). The authors of the SKAT suggest weights that are a function of a Beta(1,25) density at the pooled sample MAF to increase the contributions of rare variants to the overall sum [Wu et al., 2011]. With \( \omega_j = 1 \) for all \( j \), \( Q_{SKAT} \) is equivalent to
the C-alpha statistic [Wu et al., 2011]. With $\omega_j = \widehat{\text{Var}}_{H_0}(U_j)^{-1}$, where $\widehat{\text{Var}}_{H_0}(U_j)$ is the estimated null variance of the single-variant Cochran-Armitage trend score statistic from (2-1), $Q_{SKAT}$ is equivalent to the sum $T$ statistic (see Subsection 2.5.3). With missing genotypes and $\omega_j = \widehat{\text{Var}}_{H_0}(U_j)^{-1}$, $Q_{SKAT}$ remains equivalent to sum $T$ when the single-variant SKAT score statistics and $\omega_j$ are calculated using all available genotype data at each variant (see Subsection 2.5.3). Thus, the performance of the CA sum test should also provide insight into newer tests that achieve their robustness to neutral and protective variants by performing inference on weighted sums of nonnegative single-variant test statistics.

The implementation of the CMC test [Li and Leal, 2008] collapsed rare variants having overall sample MAF $\leq 0.01$ into an indicator variable equaling 1 if any minor alleles were present and zero otherwise. Common variants that were not collapsed were coded as 0, 1, or 2 minor alleles, and the means of the random vectors comprising the rare variant indicator and common variant minor allele counts were compared between cases and controls using Hotelling’s $T^2$ test.

One issue not considered in the paper proposing the CMC test is that LD among common variants can induce linear dependency in this random vector, which leads to a singular covariance matrix. However, calculating Hotelling’s $T^2$ statistic with any generalized inverse is equivalent to calculating the statistic with a standard inverse on a full-rank subset of linearly independent vector elements (see Subsection 2.5.4). Goodnight [Goodnight, 1979] provides an algorithm for automatically calculating a $g^2$ generalized inverse and the dimension of the full-rank subset without any prior
knowledge of the full-rank subset. The algorithm involves applying the G2SWEEP operator once to each of the columns of the covariance matrix in succession. This operator zeros the rows and columns of the covariance matrix that are numerically linearly dependent. The effective number of linearly independent vector elements, \( v \), is then automatically obtained by subtracting the number of columns that are zeroed from the total number of columns in the covariance matrix. The p-value is then calculated using the \( F_{v, N-v-1} \) approximation to the distribution of the appropriately scaled Hotelling’s \( T^2 \) statistic calculated using the \( g^2 \) generalized inverse of the covariance matrix (see Subsection 2.5.4).

Only individuals with complete genotype data at common variants could be used in calculating Hotelling’s \( T^2 \). Provided that genotype data were complete at all common variants, individuals with missing genotype data at rare variants could be used if at least one minor allele was present for a variant with a non-missing genotype because the coding of the rare variant indicator would be 1 regardless of the other variant genotypes. However, if an individual did not have any minor alleles at any variants with non-missing genotypes, the coding of the rare variant indicator was ambiguous because it would depend on the values of the unobserved genotypes. Therefore, such individuals also had to be excluded from calculating Hotelling’s \( T^2 \). With large numbers of exclusions, the \( F \) test for Hotelling’s \( T^2 \) often could not be performed due to insufficient effective denominator degrees of freedom (ddf) or was performed with only a very small number of effective ddf. Only results from \( F \) tests with effective ddf\( >4 \) were included in the Type I error and power estimates because (1) testing indicated that algebraically identical generalized inverses could yield different numerical results with effective ddf\( \leq 4 \) and (2)
the expectation and variance of the \( F \) distribution only exist for \( ddf>2 \) and \( ddf>4 \), respectively [Casella and Berger, 2002].

The implementation of the WSS test [Madsen and Browning, 2009] followed the description in the original paper with four modifications. First, midranks were used to break ties in genetic scores when calculating the case rank-sum statistic, \( W \). Second, a two-sided p-value was used. A one-sided p-value will only be well-powered for a deviation from the null in which the cumulative number of minor alleles at lower-frequency variants is higher in cases than controls. However, any departure from the null of equal genotype frequencies in cases and controls at the locus is of interest in association testing, which is why the BC-CA, CA max, CA sum, and CMC tests all use two-sided p-values. Therefore, one would also want to be able to detect deviations in which controls have a higher cumulative number of minor alleles at lower-frequency variants, which is not possible with a one-sided WSS p-value. Such deviations could arise in plausible situations, such as one in which the minor allele of a rare risk variant with a strong effect appears exclusively on a haplotype with few other minor alleles. Third, the two-sided p-value was estimated directly from the permutation distribution of \( W \). Letting \( W_t \) denote the value in permutation \( t \), \( W_{obs} \) denote the observed value in the sample, and \( \overline{W} \) denote the mean of \( W \) over all \( m \) permutations, the two-sided p-value was calculated from \( m=10,000 \) permutations as [Davison and Hinkley, 1997; Ernst, 2004]:

\[
\hat{P}_{WSS} = \frac{\sum_{t=1}^{m} I[|W_t - \overline{W}| \geq |W_{obs} - \overline{W}|] + 1}{m + 1}
\]

Finally, missing single-variant genotypes, which were not considered in the paper proposing the WSS method, were not used in estimation of the MAF in controls and were
assigned values of 0 so as not to contribute to the WSS in an individual. This procedure is
equivalent to calculating the genetic score over only nonmissing genotypes in each
individual.

2.4.6 Monte Carlo Simulation Results

The variants in the populations of haplotypes simulated based on a coalescent
model closely resembled those in the actual sequence data in Section 2.2. First, the rates
of variants per kb were compatible when the sampling process that generated the actual
sequence data was taken into account. Each simulated population had an average of 981
variant sites (range: 805 – 1193) over the 100 kb locus, or approximately 10 variants per
kb. While this rate was somewhat higher than the observed rate of 4 variants per kb
scanned in the actual sequence data, it was not inconsistent with this observation because
fewer variants are expected to be observed in any small sample from a large population.
In fact, an average of only 747 variants, or 7.5 per kb, appeared in samples of 500
individuals drawn from these haplotype populations under a null disease model with
complete genotype data. A further reduction in the number of variants per kb would be
expected in a sample of the same size as the one considered in Section 2.2, which was
about one-third the size of the smallest simulated samples.

Second, the variant MAF and pairwise LD distributions across the populations of
simulated haplotypes (Figure 2.5) closely resembled those across the six candidate genes
for dilated cardiomyopathy (Figure 2.1). The only noticeable difference between the
MAF distributions occurred in the lower quantiles because the sample MAF could not
fall below 1/368=0.002717 in the actual sequence data. The distributions of pairwise LD,
measured by the correlation coefficient, were also similar, with a strong resemblance
between the histograms and a close correspondence between the quantiles for actual and simulated data. These results suggest that variants in the average simulated haplotype population had similar MAF and LD spectra to variants in the average resequenced dilated cardiomyopathy candidate gene.

Under a null model with no risk variants, all techniques controlled type I error at the nominal level (Figure 2.6). The exact binomial 95% confidence intervals [Leemis and Trivedi, 1996] for the CA max, CA sum, and WSS test rejection rates all contained the nominal \( \alpha \) level under nearly all conditions, reflecting an observed distribution of p-values extremely close to the uniform expected under the null hypothesis. The CMC test based on Hotelling’s \( T^2 \) was often conservative in complete data, with 95% upper confidence limits below the nominal \( \alpha \) level. This result agreed with that of Li and Leal [Li and Leal, 2008], who observed conservatism increasing with the number of variants

Figure 2.5: MAF and pairwise LD distributions in simulated sequence data
Distributions of MAFs (Panel A) and pairwise LD (Panel B) for biallelic variants in 1,000 populations of 10,000 simulated haplotypes each at a 100 kb locus. Pairwise LD was measured by the within-gene pairwise correlation coefficient (\( r \)) between major/minor alleles. Because it was computationally infeasible to summarize hundreds of millions of pairwise LD values, a 0.1% simple random sample of these values was taken from each haplotype population. This sampling procedure was repeated several times with similar results. The vertical dashed line in Panel B indicates \( r = 0 \).
Figure 2.6: Simulated type I error rate comparison
Monte Carlo estimates of rejection rates for each association testing procedure based on 1,000 samples from a null disease model with no risk variants. Estimates are reported by call rate, nominal $\alpha$ level, and sample size ($N$). Error bars represent exact binomial 95% confidence intervals for the rejection rate, and dashed horizontal lines are included at the nominal $\alpha$ level. The CMC could not be performed at a call rate of 95% because no individual had complete genotype data in any sample; at a call rate of 99.5%, CMC results with $F_{ddf}>4$ were available in 23, 619, and 992 samples for $N=500$, 1,000, and 2,000, respectively.
Figure 2.7: Simulated power comparison for rare risk variants (MAF<0.005; OR=3)
Monte Carlo estimates of rejection rates for each association testing procedure based on 1,000 samples from a disease model with 50 rare risk variants (MAF<0.005; OR=3), which represent ~5% of all variants in the locus in the average population. Estimates are reported by call rate, nominal α level, and sample size (N). Error bars represent exact binomial 95% confidence intervals for the rejection rate. The CMC could not be performed at a call rate of 95% because no individual had complete genotype data in any sample; at a call rate of 99.5%, CMC results with F ddf>4 were available in 22, 596, and 991 samples for N=500, 1,000, and 2,000, respectively.
Monte Carlo estimates of rejection rates for each association testing procedure based on 1,000 samples from a disease model with 50 rare risk variants (MAF<0.01; OR=2), which represent ~5% of all variants in the locus in the average population. Estimates are reported by call rate, nominal $\alpha$ level, and sample size ($N$). Error bars represent exact binomial 95% confidence intervals for the rejection rate. The CMC could not be performed at a call rate of 95% because no individual had complete genotype data in any sample; at a call rate of 99.5%, CMC results with $F_{ddf}>4$ were available in 18, 573, and 986 samples for $N=500, 1,000, and 2,000$, respectively.
Monte Carlo estimates of rejection rates for each association testing procedure based on 1,000 samples from a disease model with 50 total risk variants, which represent ~5% of all variants in the locus in the average population, randomly allocated between rare variants (MAF<0.01; OR=2), low-frequency variants (0.01≤MAF<0.05; OR=1.5), and common variants (0.05≤MAF<0.10; OR=1.2). Estimates are reported by call rate, nominal \( \alpha \) level, and sample size \( N \). Error bars represent exact binomial 95% confidence intervals for the rejection rate. The CMC could not be performed at a call rate of 95% because no individual had complete genotype data in any sample; at a call rate of 99.5%, CMC results with \( F \) df>4 were available in 15, 564, and 989 samples for \( N=500, 1,000, \) and 2,000, respectively.
analyzed when applying Hotelling’s $T^2$ to a random vector containing between 5 and 20 rare variants in balanced case-control samples of sizes 500 and 2,000. In the null simulations presented here, Hotelling’s $T^2$ was applied to a much larger vector containing at least 110 effectively linearly independent elements in all data sets with complete genotype data. Finally, the BC-CA test nearly always had the lowest Type I error of all tests considered, reflecting conservatism due to failure to account for LD-induced correlations between single-variant test statistics.

The type I error results for the CMC test with missing genotypes also reflect the substantial loss of sample information resulting from having to exclude all individuals with incomplete genotype data at common variants. With a call rate of 95%, no type I error rate could be estimated because no individual had complete data in any of the 1,000 samples and the CMC test could not be performed. With a call rate of 99.5%, about 92% of individuals in the average sample were unusable due to missing genotype data for each sample size. For this reason, only 23 of the 1,000 samples had reliable $F$ tests with $\text{ddf}>4$ for $N=500$, and the type I error rate estimates for both $\alpha$ levels had wide 95% confidence intervals.

Under a disease model with 50 rare risk variants (MAF<0.005; OR=3), which represent ~5% of all variants in the locus in the average haplotype population, the CA max test had higher power than the CMC and WSS tests under all conditions (Figure 2.7). It also had power comparable to or higher than the CA sum test, which is equivalent to a permutation-based SKAT under an additive genetic model without covariates using the inverse of the estimated null variances of the score statistics as weights. As expected, the CA max test substantially outperformed the BC-CA test, which does not account for
LD-induced correlations between test statistics. As the sample size grew, the power of the CMC test with complete data approached that of the CA max test. With missing data, however, the CMC test generally had the lowest power due to the substantial loss of sample when it could even be performed. The CA sum test was more powerful than the CMC test under most conditions, but it began to lag the CMC in complete data for $N \geq 1,000$. The CA sum test was always more powerful than the WSS test. Although the CA sum test was more powerful the BC-CA test under all scenarios with $\alpha = 0.05$, its power deteriorated to below that of the BC-CA test in larger sample sizes with $\alpha = 0.01$.

Although the WSS test was more powerful than the BC-CA test when $N = 500$, it began to lag the BC-CA test for $N \geq 1,000$, sometimes substantially. This observation suggests that, when summing over minor alleles to reduce the number of tests performed, the power gain from reducing the multiple-testing penalty was rapidly outweighed by the power loss due to increased noise and masking as the sample size grew. The WSS test also had lower power than the CMC test in most scenarios with complete data. Because the CMC test collapses over only variants with MAF $\leq 0.01$ and analyzes common variants in a manner not subject to masking, it may perform better than the WSS test, which sums over all variants.

Results were similar under a disease model with 50 rare risk variants (MAF $\leq 0.01$; OR=2) (Figure 2.8). The CA max test had power greater than the CA sum, CMC, WSS, and BC-CA tests under all scenarios. The CA sum test continued to have higher power than the WSS test under all scenarios and was also more powerful than the CMC test for all conditions except $N = 2,000$ with complete data. The CA sum test was more powerful
than the BC-CA test under all conditions other than \( N=2,000 \) and \( \alpha=0.01 \). The WSS test also exhibited a similar pattern of performance relative to the CMC and BC-CA tests.

In the disease model with 50 total risk variants randomly allocated between rare variants (\( \text{MAF}<0.01; \text{OR}=2 \)), low-frequency variants (\( 0.01<\text{MAF}<0.05; \text{OR}=1.5 \)), and common variants (\( 0.05<\text{MAF}<0.10; \text{OR}=1.2 \)), the CA max and CA sum tests were both more powerful than the CMC and WSS tests under nearly all conditions (Figure 2.9). Under this disease model, the CA sum test, rather than the CA max test, had the highest power under all conditions. The CA max and CA sum tests also both had greater power than the BC-CA test in all scenarios. With \( N=2,000 \) and complete data, the CMC test had comparable power to the CA max test but was still less powerful than the CA sum test. The WSS test showed the same pattern of having higher power than the CMC and BC-CA tests for \( N=500 \) but beginning to lag these tests for \( N \geq 1,000 \).

2.5 Detailed Derivations

2.5.1 Particular Choices of Scores in the Generalized Cochran-Armitage Chi-Square Statistic Yield Familiar Tests

This subsection provides proofs that particular choices of scores in the generalized Cochran-Armitage test statistic defined by equation (2-1) yield a single-variant Cochran-Armitage trend test, a Pearson chi-square test for independence between the presence/absence of minor alleles at the locus and case status (collapsing), or a Z test for differences in the mean minor allele count between cases and controls (summing). For the single-variant Cochran-Armitage trend test, let \( X_k = G_{jk} \), \( r_l \) be the count of cases with \( G_{jk}=l \) (\( l=0, 1, \text{ or } 2 \)), \( s_l \) be the count of controls with \( G_{jk}=l \), and \( n_l=r_l+s_l \). The formulas for \( U \) and its estimated null variance in (2-1) can be expressed as:
These expressions are identical to those for the single-variant Cochran-Armitage trend test in Freidlin et al. [Freidlin et al., 2002]. The square of the test statistic presented in this reference has the same asymptotic $\chi^2$ null distribution as the generalized Cochran-Armitage chi-square statistic defined in (2-1).

For the collapsing test, one could use the Pearson chi-square test for independence between the presence/absence of minor alleles at the locus and case status. Without loss of generality, let $G_1$ be the multi-variant genotype where $G_{j1} = 0$ for all $j$ (i.e., that contains no minor alleles), $X_k = 1$ for $k=1$, and $X_k = 0$ for $k>1$. Then:

$$ U = \sum_k G_{jk} \left( \frac{S}{N} r_k - \frac{R}{N} s_k \right) = \sum_k \sum_{l=0}^2 I \left[ G_{jk} = l \right] \left[ \frac{S}{N} r_k - \frac{R}{N} s_k \right] $$

$$ = \sum_{l=0}^2 I \left( \frac{S}{N} \sum_k I \left[ G_{jk} = l \right] r_k - \frac{R}{N} \sum_k I \left[ G_{jk} = l \right] s_k \right) = \sum_{l=0}^2 I \left( \frac{S}{N} r_l - \frac{R}{N} s_l \right) $$

$$ \overline{\text{Var}}_{H_0} (U) = \frac{RS}{N^3} \left( N \sum_k G_{jk}^2 n_k - \left( \sum_k G_{jk} n_k \right)^2 \right) $$

$$ = \frac{RS}{N^3} \left( N \sum_{l=0}^2 \sum_k I \left[ G_{jk} = l \right] r_k^2 n_k - \left( \sum_k \sum_{l=0}^2 I \left[ G_{jk} = l \right] r_k n_k \right)^2 \right) $$

$$ = \frac{RS}{N^3} \left( N \sum_{l=0}^2 \sum_k I \left[ G_{jk} = l \right] n_k - \left( \sum_{l=0}^2 I \left[ G_{jk} = l \right] n_k \right)^2 \right) $$

$$ = \frac{RS}{N^3} \left( N \sum_{l=0}^2 n_l - \left( \sum_{l=0}^2 n_l \right)^2 \right) $$
\[ T = \frac{U^2}{\text{Var}_{H_0}(U)} = \left( \frac{S}{N} r_i - \frac{R}{N} s_i \right)^2 \] 
\[ = N \left( (s_1 + s_{-1}) r_i - (r_1 + r_{-1}) s_i \right)^2 \] 
\[ = N \left( (s_{-1} r_i - r_{-1} s_i \right)^2 \] 
where \( r_{-1} \) and \( s_{-1} \) denote the numbers of cases and controls, respectively, with genotypes other than \( G_1 \). Thus, \( T \) with this choice of scores exactly equals the desired Pearson chi-square statistic [Sokal and Rohlf, 1995], which has the same asymptotic \( \chi^2 \) null distribution as the generalized Cochran-Armitage chi-square statistic defined in (2-1).

For the summing test, one could compare the mean minor allele count over the locus between cases and controls using a \( Z \) test. Suppose that \( X_k = \sum_j G_{jk} \), the number of minor alleles summed across variants in the multi-variant genotype \( G_k \), is used as a score. Then:

\[ U = \sum_k X_k \left( \frac{S}{N} r_k - \frac{R}{N} s_k \right) = \frac{S}{N} \sum_k X_k r_k - \frac{R}{N} \sum_k X_k s_k \]
\[ = \frac{RS}{N} \left( \frac{1}{R} \sum_k X_k r_k - \frac{1}{S} \sum_k X_k s_k \right) = \frac{RS}{N} (\bar{X}_d - \bar{X}_c) \]
\[ \text{Var}_{H_0}(U) = \frac{RS}{N^3} \left( N \sum_k X_k^2 n_k - \left( \sum_k X_k n_k \right)^2 \right) \]
\[ = \frac{RS}{N^2} \left( \sum_k X_k^2 n_k - N \left( \frac{1}{N} \sum_k X_k n_k \right)^2 \right) \]
\[ = \frac{RS}{N} \text{Var}_{H_0}(X) \]
\[ T = U^2 / \text{Var}_{H_0}(U) = \frac{R^2 S^2}{N^2} (\bar{X}_A - \bar{X}_C)^2 / \left( \frac{RS}{N} \text{Var}_{H_0}(X) \right) \]

Thus, \( T \) is the square of a \( Z \) statistic for the difference in mean minor allele counts between cases and controls using a variance estimate that is consistent under the null hypothesis. In large samples, this \( Z \) statistic has a standard normal distribution under the null hypothesis, meaning that its square has the same asymptotic \( \chi^2 \) null distribution as the generalized Cochran-Armitage chi-square statistic defined in (2-1).

### 2.5.2 A Lower Bound for the Locus-Wide Power Function of the BC-CA Test

This subsection establishes that a lower bound for the locus-wide power function of the BC-CA test under the simple model presented in Section 2.3 is the power of the Bonferroni-corrected Cochran-Armitage trend test for the risk variant alone. The BC-CA test will reject the null for the locus if the maximum test statistic in the locus exceeds the Bonferroni-corrected critical value. Thus, rejection by the BC-CA test comprises 3 disjoint events:

1) The Cochran-Armitage trend chi-square statistic for variant 1 (neutral) is greater than the Bonferroni-corrected critical value \( (T_1 > \chi^2_{1,1 - \alpha/2}) \), and the Cochran-Armitage trend chi-square statistic for variant 2 (risk) is not \( (T_2 \leq \chi^2_{1,1 - \alpha/2}) \).

2) \( T_1 \leq \chi^2_{1,1 - \alpha/2} \), and \( T_2 > \chi^2_{1,1 - \alpha/2} \).

3) \( T_1 > \chi^2_{1,1 - \alpha/2} \), and \( T_2 > \chi^2_{1,1 - \alpha/2} \).
The locus-wide power function corresponds to the probabilities of these 3 disjoint events under the alternative hypothesis:
\[
\beta_{BC-CA} \left( \alpha, N, \mu_{1,a}, \sigma^2_{1,a}, \sigma^2_{1,a}, \mu_{2,a}, \sigma^2_{2,a}, \sigma^2_{2,a} \right) = P_{H_a} \left( T_1 > \chi^2_{1,1/2}, T_2 \leq \chi^2_{1,1/2} \right) + P_{H_a} \left( T_1 \leq \chi^2_{1,1/2}, T_2 > \chi^2_{1,1/2} \right) + P_{H_a} \left( T_1 > \chi^2_{1,1/2}, T_2 \leq \chi^2_{1,1/2} \right) + P_{H_a} \left( T_2 > \chi^2_{1,1/2} \right)
\]
where the second equality follows from the fact that summing over both possible events with respect to neutral variant 1 in the second and third terms yields the marginal probability of a rejection at the risk variant 2. The first term is quite difficult to evaluate without assumptions regarding the joint distribution of the single-variant test statistics, but it is always greater than or equal to zero because it is a probability. Thus, in the presence of LD:
\[
\beta_{BC-CA} \left( \alpha, N, \mu_{1,a}, \sigma^2_{1,a}, \sigma^2_{1,a}, \mu_{2,a}, \sigma^2_{2,a}, \sigma^2_{2,a} \right) = P_{H_a} \left( T_2 > \chi^2_{1,1/2} \right) = \beta \left( \frac{\alpha}{2}, N, \mu_{2,a}, \sigma^2_{2,a}, \sigma^2_{2,a} \right)
\]

2.5.3 Equivalence of the SKAT Statistic with Inverse Variance Weights to the Sum T Statistic

Following the development in Wu et al. [Wu et al., 2011], let \( y_i \) be a 1/0 case/control indicator variable and \( G_{ij} \) be the number of minor alleles (0, 1, or 2) for individual \( i \) at variant \( j \). Consider a logistic regression model for a single variant in a case-control study without additional covariates:
\[
\logit P \left( y_i = 1 \right) = \alpha_0 + \beta_j G_{ij}
\]
The SKAT statistic can be expressed as \( Q_{SKAT} = \sum_{j=1}^{\nu} \omega_j U_j^2 \), where \( U_j = G_j^T (y - \hat{\mu}_o) \) is the single-variant score statistic for testing the null hypothesis that \( \beta_j = 0 \) for variant \( j \),
\( \omega_j \) is a pre-specified weight for variant \( j \), and \( \hat{\mathbf{u}}_\theta = \expit(\hat{\alpha}_\theta) \mathbf{1}_{N \times 1} = \frac{\mathbf{R}}{N} \mathbf{1}_{N \times 1} \) for all \( j \) [Wu et al., 2011]. With \( \mathbf{G}_k = [G_{jk}] \), \( k = 1, 2, \) or \( 3 \), \( U_j \) can also be expressed as:

\[
U_j = \mathbf{G}_j^T \left( \mathbf{y} - \frac{\mathbf{R}}{N} \mathbf{1}_{N \times 1} \right) = \sum_{i=1}^N \mathbf{G}_j \left( y_i - \frac{\mathbf{R}}{N} \right)
\]

\[
= \sum_{k=1}^3 \sum_{i=1}^N I \left[ G_{ij} = G_{jk} \right] \left[ G_{jk} \left( y_i - \frac{\mathbf{R}}{N} \right) \right]
\]

\[
= \sum_{k=1}^3 G_{jk} \left( \sum_{i=1}^N I \left[ G_{ij} = G_{jk} \right] y_i - \sum_{i=1}^N I \left[ G_{ij} = G_{jk} \right] \frac{\mathbf{R}}{N} \right)
\]

\[
= \sum_{k=1}^3 G_{jk} \left( r_k - \frac{\mathbf{R}}{N} n_k \right) = \sum_{k=1}^3 G_{jk} \left( \frac{\mathbf{R} + \mathbf{S}}{N} r_k - \frac{\mathbf{R}}{N} \left( r_k + s_k \right) \right)
\]

\[
= \sum_{k=1}^3 G_{jk} \left( \frac{\mathbf{S}}{N} r_k - \frac{\mathbf{R}}{N} s_k \right)
\]

Thus, \( U_j \) is identical to \( U \) in (2-1) with \( \mathbf{G}_k = [G_{jk}] \), \( k = 1, 2, \) or \( 3 \), and \( X_k = G_{jk} \).

Now suppose that \( \omega_j = \sqrt{\text{Var}_{H_0} \left( U_j \right)^{-1}} \), where \( \sqrt{\text{Var}_{H_0} \left( U_j \right)} \) is the estimated null variance of the single-variant Cochran-Armitage trend score statistic from (2-1). This choice of weights increases the contributions of rare variants to the overall SKAT statistic in a principled manner by weighting sum components so that they have the same marginal asymptotic distribution under the null hypothesis. With these weights:

\[
\mathbf{Q}_{SKAT} = \sum_{j=1}^v \omega_j U_j^2 = \sum_{j=1}^v \frac{U_j^2}{\sqrt{\text{Var}_{H_0} \left( U_j \right)}} = \sum_{j=1}^v T_j
\]

\( T_j \) is clearly the single-variant Cochran-Armitage trend chi-square statistic for variant \( j \) from (2-1). Thus, without covariates, with complete genotype data, and with \( \omega_j = \sqrt{\text{Var}_{H_0} \left( U_j \right)^{-1}} \), \( \mathbf{Q}_{SKAT} \) is equivalent to the sum \( T \) statistic.
This equivalence continues to hold with missing genotypes provided that the single-variant SKAT score statistics and $\omega_j$ are calculated using all available genotype data at each variant. For an arbitrary variant $j$, let $O_j$ be a vector of length $N$ containing 1 for individuals with available genotype data and 0 otherwise. Also let a superscript $O$ on a previously defined quantity denote that it was calculated in all individuals with available genotype data at variant $j$. The single-variant score statistic for testing the null hypothesis that $\beta_j = 0$ in the $N_j^O$ individuals with available genotype data at variant $j$ is

$$U_j^O = (O_j^\circ G_j)^T (y - \hat{\mu}_{0j}^O),$$

where $\hat{\mu}_{0j}^O = \expit(\hat{\alpha}_{0j}^O) I_{N^O} = \frac{R_{j}^{O}}{N_j^O} I_{N^O}$. $U_j^O$ can also be expressed as:

$$U_j^O = (O_j^\circ G_j)^T \left( y - \frac{R_{j}^{O}}{N_j^O} I_{N^O} \right) = \sum_{i=1}^{N} O_y G_y \left( y_i - \frac{R_{j}^{O}}{N_j^O} \right)$$

$$= \sum_{k=1}^{3} G_{jk} \left( \sum_{i=1}^{N} O_y I \left[ G_y = G_{jk} \right] y_i - \sum_{i=1}^{N} O_y I \left[ G_y = G_{jk} \right] R_{j}^{O} \right)$$

$$= \sum_{k=1}^{3} G_{jk} \left( r_k^O - \frac{R_{j}^{O}}{N_j^O} n_k^O \right) = \sum_{k=1}^{3} G_{jk} \left( \frac{S_k^O}{N_j^O} r_k^O - \frac{R_{j}^{O}}{N_j^O} \right)$$

$U_j^O$ is therefore identical to $U$ in (2.1) calculated in the $N_j^O$ individuals with available genotype data at variant $j$ using $G_k = \left[ G_{jk} \right]$, $k=1, 2, \text{ or } 3$, and $X_k = G_{jk}$. Denote $\widehat{\text{Var}}_{H_0}(U_j)$ calculated in these individuals by $\widehat{\text{Var}}_{H_0}(U_j^O)$ and set $\omega_j = \widehat{\text{Var}}_{H_0}(U_j^O)^{-1}$. It follows that:
\[
Q_{SKAT} = \sum_{j=1}^{V} \omega_j \left( U_{j}^O \right)^2 = \sum_{j=1}^{V} \frac{\left( U_{j}^O \right)^2}{\text{Var}_H(U_{j}^O)} = \sum_{j=1}^{V} T_j^O
\]

\( T_j^O \) is the single-variant Cochran-Armitage trend chi-square statistic calculated using all available genotypes at variant \( j \). Thus, \( Q_{SKAT} \) remains equivalent to sum \( T \) when the single-variant SKAT score statistics and \( \omega_j \) are calculated using all available genotype data at each variant.

### 2.5.4 Performing Hotelling’s \( T^2 \) Test on a Maximal Linearly Independent Subset of Vector Elements by Using a Generalized Inverse

Let \( \mathbf{x}_i \) and \( \mathbf{y}_j \) be column vectors of dimension \( u \) indexed by \( i=1,\ldots,n_x \) and \( j=1,\ldots,n_y \). In what follows, any quantity with a bar over it denotes the mean over \( i \) or \( j \), as appropriate, and a superscript \( T \) indicates the transpose. The pooled covariance matrix \( \mathbf{S} \) is:

\[
\mathbf{S} = (n_x + n_y - 2)^{-1} \left[ \sum_{i=1}^{n_x} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T + \sum_{j=1}^{n_y} (\mathbf{y}_j - \bar{\mathbf{y}})(\mathbf{y}_j - \bar{\mathbf{y}})^T \right]
\]

Hotelling’s \( T^2 \) test statistic using a generalized inverse of \( \mathbf{S}, \mathbf{S}^{-} \), is defined as:

\[
T^2 = \frac{n_x n_y}{n_x + n_y} (\bar{\mathbf{x}} - \bar{\mathbf{y}})^T \mathbf{S}^{-} (\bar{\mathbf{x}} - \bar{\mathbf{y}})
\]

(2-4)

Suppose that some elements of the \( \mathbf{x}_i \) and \( \mathbf{y}_j \) vectors are perfectly predicted by linear combinations of the others, such as if genotypes at a particular variant can be perfectly predicted based on linear combinations of the genotypes at other variants due to LD. In that case, there exists a permutation matrix \( \mathbf{P} \) that rearranges the \( \mathbf{x}_i \) and \( \mathbf{y}_j \) vectors such that the first \( v \) elements contain all the linearly independent elements of the original
vectors and the remaining $u - v$ elements contain all the linearly dependent elements.

Mathematically,

\[
P_{x_i} = \begin{bmatrix} x_{ii} \\ x_{di} \end{bmatrix} = \begin{bmatrix} x_{ii} \\ Lx_{ii} \end{bmatrix}
\]

\[
P_{y_j} = \begin{bmatrix} y_{ij} \\ y_{dj} \end{bmatrix} = \begin{bmatrix} y_{ij} \\ Ly_{ij} \end{bmatrix}
\]

where the subscript $I$ refers to the $v$ linearly independent elements, the subscript $D$ refers to the $u - v$ linearly dependent elements, and $L$ is an arbitrary $(u - v) \times v$ linear transformation defining the linear dependency of the $D$ elements on the $I$ elements.

Applying this permutation and its transpose to the covariance matrix yields:

\[
\mathbf{PSP}^T = (n_x + n_y - 2)^{-1} \left[ \sum_{i=1}^{n_x} \mathbf{P}(x_i - \bar{x})(x_i - \bar{x})^T \mathbf{P}^T + \sum_{j=1}^{n_y} \mathbf{P}(y_j - \bar{y})(y_j - \bar{y})^T \mathbf{P}^T \right]
\]

\[
= (n_x + n_y - 2)^{-1} \left[ \sum_{i=1}^{n_x} \left[ L(x_{ii} - \bar{x}) \right] \left[ L(x_{ii} - \bar{x}) \right]^T + \sum_{j=1}^{n_y} \left[ L(y_{ij} - \bar{y}) \right] \left[ L(y_{ij} - \bar{y}) \right]^T \right]
\]

\[
= (n_x + n_y - 2)^{-1} \left[ \sum_{i=1}^{n_x} \left( x_{ii} - \bar{x}_i \right) \left( x_{ii} - \bar{x}_i \right)^T + \sum_{j=1}^{n_y} \left( y_{ij} - \bar{y}_j \right) \left( y_{ij} - \bar{y}_j \right)^T \right] + \sum_{j=1}^{n_y} \left[ L(y_{ij} - \bar{y}_j) \right] \left[ L(y_{ij} - \bar{y}_j) \right]^T
\]

\[
= \left[ \mathbf{S}_I \quad \mathbf{S}_i \mathbf{L}^T \right] = \left[ \mathbf{S}_v \quad \mathbf{S}_u \right] = \left[ \mathbf{S}_V \right]
\]

where $\mathbf{S}_I$ is the $v \times v$ covariance matrix for the $I$ elements.
By definition, $S_U = S_V L^T$, meaning that all columns of $S$ moved to $S_U$ are linear combinations of other columns of $S$ moved to $S_V$. Therefore, $S_U$ cannot contain any linearly independent column of $S$, and $S_V$ contains all the linearly independent columns of $S$. Now, let $X_{IC}$ be the $n_x \times v$ matrix obtained by stacking $(x_{ii} - \bar{x}_i)^T$ in the rows. This matrix has rank $v$ because the columns corresponding to the vector elements are linearly independent by definition. The sum of squares and cross products matrix for $x_{ii}$ can then be written as:

$$\sum_{i=1}^{n_i} (x_{ii} - \bar{x}_i)(x_{ii} - \bar{x}_i)^T = X_{IC}^T X_{IC} = X_{IC}^T I_{n_x \times n_x} X_{IC}$$

Because $I_{n_x \times n_x}$ is positive definite and $X_{IC}$ has rank $v$, $X_{IC}^T I_{n_x \times n_x} X_{IC}$ is also positive definite (Theorem 14.2.9 of Harville [Harville, 1997]). The same reasoning follows for the sum of squares and cross products matrix for $y_{ii}$. Now, $S_I$ can be written as the sum of two positive definite matrices:

$$S_I = (n_x + n_y - 2)^{-1} \left[ X_{IC}^T I_{n_x \times n_x} X_{IC} + Y_{IC}^T I_{n_y \times n_y} Y_{IC} \right]$$

Because the sum of two positive definite matrices is also positive definite (Lemma 14.2.4 of Harville [Harville, 1997]) and any positive definite matrix is nonsingular (Lemma 14.2.8 of Harville [Harville, 1997]), $S_I$ has full rank. Because $S_V$ contains all the linearly independent columns of $S$, a subset of the columns in $S_V$ forms a basis for the column space of $S$. An arbitrary linearly dependent column of $S$ is therefore a linear combination of a subset of columns moved to $S_V$. If a linearly dependent column in $S$ had been moved to $S_V$, it would also have to be linearly dependent on some other columns in $S_V$, which would contradict the fact that $S_I$ has full rank. Thus, $S_V$ cannot contain any linearly dependent column of $S$, meaning that $S_V$ contains only linearly independent columns of
Therefore, the first \(v\) columns of \(\text{PSP}^T\) contain all the linearly independent columns of \(S\) and only these columns, and the same result for the first \(v\) rows of \(\text{PSP}^T\) immediately follows from symmetry.

To summarize, \(S\) is a \(u \times u\) symmetric matrix of rank \(v < u\), and \(P\) is permutation matrix that generates a partitioned matrix \(\text{PSP}^T\) in which the first \(v\) columns and rows are the linearly independent columns and rows of \(S\). According to Theorem 9.2.5 of Harville [Harville, 1997], \(S^{-}\) is a generalized inverse of \(S\) if and only if:

\[
S^{-} = P^T \begin{bmatrix}
S_i^{-1} - ALS_iS_i^{-1} - S_i^{-1}S_iL^T B - S_i^{-1}S_iL^T C S_i S_i^{-1} & A \\
B & C
\end{bmatrix} P
\]

for arbitrary matrices \(A\), \(B\), and \(C\) of the appropriate dimensions. Specific types of generalized inverses, such as \(g2\) and Moore-Penrose generalized inverses, correspond to specific choices of \(A\), \(B\), and \(C\) that satisfy the additional conditions required for these types of generalized inverses.

Substituting the above expression for the generalized inverse into the Hotelling’s \(T^2\) test statistic defined in (2-4) yields:
\[ T^2 = \frac{n_x n_y}{n_x + n_y} (\bar{x} - \bar{y})^T S^{-1} (\bar{x} - \bar{y}) \]

\[ = \frac{n_x n_y}{n_x + n_y} (\bar{x} - \bar{y})^T P^T \left[ S^{-1} - A L - L^T B - L^T C L \right] \begin{bmatrix} A \\ B \\ C \end{bmatrix} P (\bar{x} - \bar{y}) \]

\[ = \frac{n_x n_y}{n_x + n_y} \left[ \bar{x}_b - \bar{y}_b \right]^T \left[ S^{-1} - A L - L^T B - L^T C L \right] \begin{bmatrix} A \\ B \\ C \end{bmatrix} \left[ \bar{x}_b - \bar{y}_b \right] \]

\[ = \frac{n_x n_y}{n_x + n_y} \left[ \bar{x}_i - \bar{y}_i \right]^T \left[ S^{-1} - A L - L^T B - L^T C L \right] \begin{bmatrix} A \\ B \\ C \end{bmatrix} L \left( \bar{x}_i - \bar{y}_i \right) \]

\[ = \frac{n_x n_y}{n_x + n_y} \left[ (\bar{x}_i - \bar{y}_i)^T (S^{-1} - A L - L^T B - L^T C L) (\bar{x}_i - \bar{y}_i) \right] \]

Thus, using any type of generalized inverse in Hotelling’s $T^2$ statistic is algebraically equivalent to calculating the statistic using the maximal linearly independent subset of $v$ vector elements. However, procedures to calculate the generalized inverse do not require any prior knowledge of which vector elements form this subset, meaning that knowledge of $P$ is not required to apply the procedure. Based on the above equivalence, inference on the test statistic defined in (2-4) can be performed using the $F$ approximation [Timm, 2002]:

\[ \frac{n_x + n_y - v - 1}{v} \frac{T^2}{n_x + n_y - 2} \sim F_{v, n_x + n_y - v - 1} \]

These algebraic equivalences may not hold numerically when there are only a few more vectors in the sample than the number of linearly independent elements (i.e., $n_x + n_y - v - 1$ is in the single digits). In such situations, small differences in the
numerical algorithms used to calculate each type of generalized inverse may lead to different test statistics due to different roundoff error properties or selection of different maximal linearly independent subsets of vector elements.


CHAPTER 3: 
VALID PERMUTATION TESTS FOR GENETIC CASE-CONTROL STUDIES 
WITH MISSING GENOTYPES 

3.1 Summary 
In this chapter, a broadly applicable method for constructing exact level-\( \alpha \) permutation tests is derived for genetic case-control studies in the presence of missing genotypes. The development begins with the specification of a nonparametric probability model for the observed data in such a study. Group-theoretic arguments are then used to establish two conditions that together guarantee an exact level-\( \alpha \) Monte Carlo permutation test for data generated under this nonparametric probability model. One of these conditions is not satisfied for the most frequently used Monte Carlo permutation test, and this test is guaranteed to be level \( \alpha \) only for missing data processes with certain characteristics. An alternative Monte Carlo permutation test, which is exact level \( \alpha \) as long as all covariates influencing the missing data process are identified and recorded, is therefore proposed. The theoretical development is supplemented with Monte Carlo simulations for a variety of test statistics and missing data processes. 

3.2 Nonparametric Probability Model 
In the subsequent exposition, an uppercase script letter (e.g., \( Z \)) denotes a set, and \( Z^n = \underbrace{Z \times \cdots \times Z}_n \) is the Cartesian product taken over \( n \) copies of the same set. 
\( I[\text{expression}] \) denotes the indicator function taking a value of 1 when the expression in brackets is true and 0 when it is false.
Let $a = I[\text{Individual is a case}]$ be a nonrandom case status indicator determined by design and $\mathbf{G} = [G_1, \ldots, G_v]$ be a random row vector of genotype scores for the $v$ variants at the locus of interest in an individual. Each element $G_j$ takes on a countable number of possible values in the space $\mathcal{G}$ and may be correlated with other elements of the random vector due to linkage disequilibrium. Following Little and Rubin, let $\mathbf{M} = [M_1, \ldots, M_v]$ represent a random row vector of missing genotype indicators, $M_j = I[G_j \text{ missing}]$ [Little and Rubin, 2002]. The complete data vector for an individual, $[\mathbf{G}, \mathbf{M}] = [G_1, \ldots, G_v, M_1, \ldots, M_v]$, is then a random row vector taking values in the sample space $\mathcal{Y} = \mathcal{G}^v \times \{0,1\}^v$. Let $f_{G,M}(\mathbf{g}, \mathbf{m}|a, \mathbf{x}, q)$ denote the conditional probability mass function (PMF) of $[\mathbf{G}, \mathbf{M}]$ over $\mathcal{Y}$, where $[\mathbf{g}, \mathbf{m}] = [g_1, \ldots, g_v, m_1, \ldots, m_v]$ represents an arbitrary vector in $\mathcal{Y}$, $\mathbf{x} = [x_1, \ldots, x_p]$ is a vector of nonrandom covariate values in the individual that affect the distribution of $[\mathbf{G}, \mathbf{M}]$, and $q$ is an arbitrary population in a space $\mathcal{Q}$ from which the individual is sampled. This PMF therefore describes both the sampling of genotype scores from the population $q$ as well as the missing data process that determines which ones are observed.

The random observed genotype vector $\mathbf{G}_{\text{obs}}$ for an individual is a coarsening of the random vector $\mathbf{G}$ that results from setting each $G_j$ to missing according to the value of $M_j$ [Little and Rubin, 2002]. One can imagine this coarsening resulting from a selection model [Little and Rubin, 2002] in which the value of $\mathbf{G}$ is determined first.
during sampling of subjects from \( q \) and then the value of \( \tilde{G}_{\text{obs}} \) is determined based on the value of \( \tilde{M} \) subsequently realized during genotype data collection. Formally, let \( h \) be an arbitrary missing genotype code not equal to any valid genotype score in \( G \) (e.g., -99 if genotype scores equal the number of minor alleles) and define:

\[
\tilde{G}_{\text{obs}} = \eta \left( \begin{bmatrix} \tilde{G}, \tilde{M} \end{bmatrix} \right) = \left[ G_I [M_1 = 0] + hI [M_1 = 1], \ldots, G_I [M_v = 0] + hI [M_v = 1] \right]
\]

Here, each \( G_{\text{obs},j} \) takes values in the sample space \( \mathcal{H} = \mathcal{G} \cup \{h\} \), each random vector \( \tilde{G}_{\text{obs}} \) takes values in the sample space \( \mathcal{Z} = \mathcal{H}^v \), and \( \eta : \mathcal{Y} \to \mathcal{Z} \). For an arbitrary vector \( \tilde{g}_{\text{obs}} = [g_{\text{obs},1}, \ldots, g_{\text{obs},v}] \in \mathcal{Z} \), the value of \( \eta^{-1} : \mathcal{Z} \to \mathcal{Y} \) is the set of one or more \([g, \tilde{m}] \in \mathcal{Y}\) that map to \( \tilde{g}_{\text{obs}} \) through \( \eta : \mathcal{Y} \to \mathcal{Z} \). This set comprises all \([g, \tilde{m}] \in \mathcal{Y}\) having the same missing data pattern and the same genotype scores for all observed genotypes, so:

\[
\eta^{-1} (\tilde{g}_{\text{obs}}) = \left\{ [g, \tilde{m}] \in \mathcal{Y} : \eta \left( \begin{bmatrix} g, \tilde{m} \end{bmatrix} \right) = \tilde{g}_{\text{obs}} \right\}
\]

\[
= \left\{ [g, \tilde{m}] \in \mathcal{Y} : g_j \in \tilde{G}, m_j = 1 \ \forall \ j \text{ with } g_{\text{obs},j} = h; \ g_j = g_{\text{obs},j}, m_j = 0 \ \forall \ j \text{ with } g_{\text{obs},j} \neq h \right\}
\]

Therefore, the conditional PMF of \( \tilde{G}_{\text{obs}} \) under \( f_{\tilde{G},\tilde{M}} \) given \( (a, \bar{x}, q) \) is:

\[
P \left( \tilde{G}_{\text{obs}} = \tilde{g}_{\text{obs}} \mid a, \bar{x}, q \right) = P \left( \eta \left( \begin{bmatrix} \tilde{G}, \tilde{M} \end{bmatrix} \right) = \tilde{g}_{\text{obs}} \mid a, \bar{x}, q \right) = P \left( [\tilde{G}, \tilde{M}] \in \eta^{-1} (\tilde{g}_{\text{obs}}) \mid a, \bar{x}, q \right)
\]

\[
= \sum_{\eta^{-1}(\tilde{g}_{\text{obs}})} f_{\tilde{G},\tilde{M}} (\tilde{g}, \tilde{m}) | a, \bar{x}, q)
\]

(3-1)

For a genetic case-control study with \( n \) total individuals, let \( a = [a_1, \ldots, a_n]^T \) be the nonrandom affection status indicator column vector containing \( a_i \) for individual \( i \) in row
$i$ and $\mathbf{x} = [\tilde{x}_i^T, \ldots, \tilde{x}_n^T]^T$ be the nonrandom covariate matrix containing the row vector $\tilde{x}_i$ for individual $i$ in row $i$. The observed genotype data for this study is represented by the random observed genotype matrix $\mathbf{G}_{\text{obs}} = [\mathbf{G}_{\text{obs},1}^T, \ldots, \mathbf{G}_{\text{obs},n}^T]^T$ containing the random row vector $\mathbf{G}_{\text{obs},j}$ for individual $i$ in row $i$. This matrix takes values $\mathbf{g}_{\text{obs}} = [g_{\text{obs},1}, \ldots, g_{\text{obs},n}]^T$ in $\mathbb{Z}^n$, where each row $i$ contains an arbitrary vector $\mathbf{g}_{\text{obs}}$ from $\mathbb{Z}$. Because individuals in a genetic case-control study are unrelated, the distribution of $\mathbf{G}_{\text{obs},j}$ conditional on $(a_j, \tilde{x}_j, q)$ does not depend on any $\mathbf{G}_{\text{obs},k|a_k}$. Therefore, each $\mathbf{G}_{\text{obs},i}$ is an independent realization from the conditional distribution in (3-1) given $(a_i, \tilde{x}_i, q)$, and the conditional PMF of $\mathbf{G}_{\text{obs}}$ under $f_{G,M}$ given $(a, \mathbf{x}, q)$ is:

$$
P(\mathbf{G}_{\text{obs}} = \mathbf{g}_{\text{obs}} | a, \mathbf{x}, q) = \prod_{i=1}^{n} P(\mathbf{G}_{\text{obs},i} = \mathbf{g}_{\text{obs},i} | a_i, \tilde{x}_i, q) = \prod_{i=1}^{n} \sum_{\eta^{-1}(\mathbf{g}_{\text{obs},i})} f_{G,M} ([\mathbf{g}, \mathbf{m}] | q, \tilde{x}_i, q)
$$

The population $q$ in a particular study is one of many possible realizations from some unknown evolutionary process. Thus, variability in $\mathbf{G}_{\text{obs}}$ arises from both statistical sampling from a particular population $q$ according to $f_{G,M}$ and genetic sampling of populations $Q$ from the unknown evolutionary process according to a PMF $f_Q$ [Weir, 1996]. The conditional PMF of $\mathbf{G}_{\text{obs}}$ given $(a, \mathbf{x})$ that incorporates both of these sources of variation is:
\[
F(g_{\text{obs}} | a, x) = \sum_{q} P(G_{\text{obs}} = g_{\text{obs}} | a, x, q) f_{\mathcal{G}}(q)
\]

\[
= \sum_{q} \prod_{i=1}^{n} \sum_{q_i^{-1}(\mathcal{D}_{\text{obs}})} f_{GM}( [g, \tilde{m}] | a, \tilde{x}, q )\]  \( f_{\mathcal{G}}(q) \)  \hspace{1cm} (3-2)

\( F \) is referred to as the data generating process.

The null hypothesis in a genetic case-control study specifies that the evolutionary process results in a population disease model that does not depend on \( g \). As a result, the sampling distribution of \( \tilde{G} \) is independent of \( a \) for all \( q \) under the null hypothesis. Given an arbitrary nominal \( \alpha \), the goal is to construct an exact level-\( \alpha \) Monte Carlo permutation test of this null hypothesis; that is, a Monte Carlo permutation test with a type I error rate of at most \( \alpha \) for all test statistics and data generating processes \( F \) consistent with this null hypothesis.

### 3.3 Group-Theoretic Construction of Exact Level-\( \alpha \) Monte Carlo Permutation Tests

Group-theoretic arguments have been used previously to solve similar testing problems in nonparametric settings [Hoeffding, 1952; Lehmann and Romano, 2005; Romano, 1989, 1990]. In this section, variations on such arguments are used to establish two conditions that together guarantee an exact level-\( \alpha \) Monte Carlo permutation test of the null hypothesis in a genetic case-control study under the nonparametric probability model given in (3-2).

Let \( \pi : \{1, \ldots, n\} \rightarrow \{1, \ldots, n\} \) be an arbitrary permutation of the integers. The composite of two arbitrary permutations \( \pi_2 \circ \pi_1 \) is the result of first applying permutation \( \pi_1 \) and then applying permutation \( \pi_2 \) to the result; that is, \( (\pi_2 \circ \pi_1)(i) = \pi_2(\pi_1(i)) \). The
composite $\pi_2 \circ \pi_1$ is also a permutation, and every $\pi$ has a unique inverse permutation $\pi^{-1}$ such that $(\pi \circ \pi^{-1})(i) = (\pi^{-1} \circ \pi)(i) = i$ for all $i$ [Robinson, 2003].

Let $u_\pi$ be the function that puts row $\pi(i)$ of its matrix argument in row $i$ of its matrix result of the same dimensions. Thus, for an arbitrary $g_{\text{obs}} \in \mathbb{Z}^n$, 

$$u_\pi \left( g_{\text{obs}} \right) = \left[ g_{\text{obs}, \pi(i)}^T, \ldots, g_{\text{obs}, \pi(n)}^T \right]^T.$$  

It is clear that $u_\pi \left( g_{\text{obs}} \right) \in \mathbb{Z}^n$ because $g_{\text{obs}, \pi(i)} \in \mathbb{Z}$ for all $i$, so $u_\pi : \mathbb{Z}^n \rightarrow \mathbb{Z}^n$. Moreover, for an arbitrary $\pi'$, row $i$ of $u_{\pi'} \left( u_\pi \left( g_{\text{obs}} \right) \right)$ contains the value of row $\pi(i)$ of $u_{\pi'} \left( g_{\text{obs}} \right)$, and row $\pi(i)$ of $u_{\pi'} \left( g_{\text{obs}} \right)$, the value of row $\pi' \left( \pi(i) \right) = (\pi' \circ \pi)(i)$ of $g_{\text{obs}}$. It follows that:

$$u_{\pi'} \left( u_\pi \left( g_{\text{obs}} \right) \right) = \left[ g_{\text{obs}, \pi'(\pi(i))}^T, \ldots, g_{\text{obs}, \pi'(\pi(n))}^T \right]^T = u_{\pi' \circ \pi} \left( g_{\text{obs}} \right)$$

Because every $\pi$ has a has a unique inverse permutation $\pi^{-1}$, it immediately follows that $u_\pi$ and $u_{\pi^{-1}}$ satisfy:

$$u_{\pi^{-1}} \left( u_\pi \left( g_{\text{obs}} \right) \right) = u_\pi \left( u_{\pi^{-1}} \left( g_{\text{obs}} \right) \right) = g_{\text{obs}}$$

Thus, $u_{\pi^{-1}}$ is the unique inverse for $u_\pi$, which implies that $u_\pi$ is a one-to-one mapping of $\mathbb{Z}^n$ onto itself [Robinson, 2003].

Constructing a permutation test of the null hypothesis in a genetic case-control study involves choosing a set $S$ of $R$ permutations as well as an arbitrary real-valued test statistic $t(G_{\text{obs}}, \mathbf{a}) = t \left( \left[ G_{\text{obs}, 1}, \mathbf{a}_1 \right], \ldots, \left[ G_{\text{obs}, n}, \mathbf{a}_n \right] \right)$. This test statistic measures the association between $G_i$ and $a_i$ using the observed genotype data and is symmetric in its
vector arguments \( \left[ \hat{\mathbf{G}}_{\text{obs},i}, \mathbf{a}_i \right] \). Without loss of generality, assume that \( t \) takes on larger values for matrices \( \mathbf{g}_{\text{obs}} \in \mathbb{Z}^n \) more inconsistent with the null hypothesis of no association between \( \hat{\mathbf{G}}_i \) and \( a_i \). The permutation p-value with particular choices of \( S \) and \( t \) is defined by:

\[
p_t^S (\mathbf{G}_{\text{obs}}, \mathbf{a}) = R^{-1} \sum_{S} I \left[ t \left( u_x (\mathbf{G}_{\text{obs}}), \mathbf{a} \right) \geq t (\mathbf{G}_{\text{obs}}, \mathbf{a}) \right]
\]

The permutation test is based on comparing the permutation p-value to an arbitrary nominal \( \alpha \in (0,1) \) and rejecting the null hypothesis with probability given by the test function:

\[
\phi_{t,\alpha}^S (\mathbf{G}_{\text{obs}}, \mathbf{a}) = \begin{cases} 
1 & p_t^S (\mathbf{G}_{\text{obs}}, \mathbf{a}) \leq \alpha \\
0 & p_t^S (\mathbf{G}_{\text{obs}}, \mathbf{a}) > \alpha 
\end{cases}
\]

The permutation test defined here is therefore a nonrandomized test that always rejects the null hypothesis for \( \mathbf{g}_{\text{obs}} \in \mathbb{Z}^n \) with permutation p-values less than or equal to \( \alpha \) and always fails to reject the null hypothesis for all other \( \mathbf{g}_{\text{obs}} \in \mathbb{Z}^n \). As the notation reflects, each choice of \( S \) and \( t \) defines a different permutation test. The goal is to choose \( S \) in a way that guarantees that the resulting permutation test is exact level \( \alpha \) (i.e., has a type I error rate of at most \( \alpha \) for all \( t \) and \( \mathbb{F} \) consistent with the null hypothesis).

If \( S \) is a group under composition, then \( \sum_S \phi_{t,\alpha}^S (u_x (\mathbf{g}_{\text{obs}}), \mathbf{a}) \leq R\alpha \) for all \( t \) and \( \mathbf{g}_{\text{obs}} \in \mathbb{Z}^n \) (see Subsection 3.7.1). This property implies that:

\[
E_x \left[ \sum_S \phi_{t,\alpha}^S (u_x (\mathbf{G}_{\text{obs}}), \mathbf{a}) \bigg| \mathbf{a}, \mathbf{x} \right] = \sum_S E_x \left[ \phi_{t,\alpha}^S (u_x (\mathbf{G}_{\text{obs}}), \mathbf{a}) \bigg| \mathbf{a}, \mathbf{x} \right] \leq R\alpha
\]
for all \( t \) and \( \mathbb{F} \). For an arbitrary \( \pi \in S \), the random matrix \( u_{\pi}(G_{obs}) \) takes on values in the space \( \{u_{\pi}(g_{obs}) : g_{obs} \in \mathbb{Z}^n\} \). Because \( u_{\pi} : \mathbb{Z}^n \rightarrow \mathbb{Z}^n \) is onto, it immediately follows that this space is simply \( \mathbb{Z}^n \). Suppose also that, for a particular \( \mathbb{F} \) consistent with the null hypothesis, the random matrix \( u_{\pi}(G_{obs}) \) has an induced PMF identical to \( \mathbb{F} \) for all \( \pi \in S \). This condition implies that:

\[
E_{\mathbb{F}} \left[ \phi_{t,\alpha}^{\delta^S}(u_{\pi}(G_{obs}), a) \bigg| a, x \right] = \sum_{z_t^n} \phi_{t,\alpha}^{\delta^S}(g_{obs}, a) \mathbb{P}_{\mathbb{F}} \left( u_{\pi}(G_{obs}) = g_{obs} \big| a, x \right) = \sum_{z_t^n} \phi_{t,\alpha}^{\delta^S}(g_{obs}, a) \mathbb{F}(g_{obs} \big| a, x) = E_{\mathbb{F}} \left[ \phi_{t,\alpha}^{\delta^S}(G_{obs}, a) \bigg| a, x \right]
\]

for all \( t \) and \( \pi \in S \). It immediately follows that \( E_{\mathbb{F}} \left[ \phi_{t,\alpha}^{\delta^S}(G_{obs}, a) \bigg| a, x \right] \leq \alpha \) for all \( t \), so the permutation test based on \( S \) has a type I error rate of at most \( \alpha \) for all \( t \) under \( \mathbb{F} \). If the random matrix \( u_{\pi}(G_{obs}) \) has an induced PMF identical to \( \mathbb{F} \) for all \( \pi \in S \) and all \( \mathbb{F} \) consistent with the null hypothesis, then the permutation test based on \( S \) is exact level \( \alpha \). Similar group invariance conditions for obtaining exact level-\( \alpha \) randomized and nonrandomized permutation tests have been previously provided by a number of authors [Hoeffding, 1952; Lehmann and Romano, 2005; Romano, 1989, 1990].

For arbitrary \( \pi \in S \) and \( g_{obs} \in \mathbb{Z}^n \), \( u_{\pi}(G_{obs}) = g_{obs} \) if and only if \( G_{obs} = u_{\pi^{-1}}(g_{obs}) \) because \( u_{\pi} \) is one-to-one and onto. The induced PMF of \( u_{\pi}(G_{obs}) \) for this \( \pi \in S \) under a particular \( \mathbb{F} \) can therefore be obtained from (3-2):
Using (3-2) and (3-3), \( u_\pi (G_{obs}) \) has an induced PMF identical to \( F \) for all \( \pi \in S \) if:

\[
\prod_{i=1}^{n} \sum_{\eta^{-1}(g_{obs}^{-1}(i))} f_{G,M}(g,\tilde{m}|a_i,\tilde{x},q) = \prod_{i=1}^{n} \sum_{\eta^{-1}(g_{obs})} f_{G,M}(g,\tilde{m}|a_i,\tilde{x},q)
\]

for all \( \pi \in S \), \( q \in Q \), and \( g_{obs} \in \mathbb{Z}^n \). This condition is referred to as exchangeability of observations [Good, 2000; Pesarin and Salmaso, 2010]. Suppose that terms in these products are now indexed by the row number in the original \( g_{obs} \), \( k \), rather than individual number \( i \). For the lefthand product, the term for \( \tilde{g}_{obs,\pi^{-1}(i)} \) has the PMF for individual \( i = \pi(\pi^{-1}(i)) \), so the term for \( \tilde{g}_{obs,k} \) must have the PMF for individual \( \pi(k) \). For the righthand product, the term for \( \tilde{g}_{obs,i} \) has the PMF for individual \( i \), so the term for \( \tilde{g}_{obs,k} \) must have the PMF for individual \( k \). Thus, \( u_\pi (G_{obs}) \) has an induced PMF identical to \( F \) for all \( \pi \in S \) if:

\[
\prod_{k=1}^{n} \sum_{\eta^{-1}(g_{obs})} f_{G,M}(g,\tilde{m}|a_{\pi(k)},\tilde{x}_{\pi(k)},q) = \prod_{k=1}^{n} \sum_{\eta^{-1}(g_{obs,k})} f_{G,M}(g,\tilde{m}|a_k,\tilde{x}_k,q)
\]

(3-4)

for all \( \pi \in S \), \( q \in Q \), and \( g_{obs} \in \mathbb{Z}^n \). Note that this condition is satisfied if the PMF for individual \( \pi(k) \) is identical to the PMF for individual \( k \) for all \( \pi \in S \), \( q \in Q \), and \( k \).

In summary, if \( S \) is a group under composition and equation (3-4) is satisfied for all \( \pi \in S \), \( q \in Q \), and \( g_{obs} \in \mathbb{Z}^n \) under a particular \( F \) consistent with the null hypothesis,
the permutation test based on $S$ is has a type I error rate of at most $\alpha$ for all $t$ under $F$.

Thus, for a permutation test based on $S$ to be exact level $\alpha$ (i.e., have a type I error rate of at most $\alpha$ for all $t$ and $F$ consistent with the null hypothesis), it is sufficient that:

(I) $S$ is a group under composition; and

(II) Equation (3-4) is satisfied for all $\pi \in S$, $q \in Q$, $g_{\text{obs}} \in Z^n$, and $F$ consistent with the null hypothesis.

In practice, the permutation p-value is generally stochastically approximated to avoid enumerating all $\pi \in S$, and affection status indicators, rather than intact observed genotype vectors, are permuted. The resulting procedure is referred to as the Monte Carlo permutation test based on $S$. Specifically, suppose that a with-replacement random sample of $W \geq (1-\alpha)/\alpha$ elements from $S$, $\Pi = (\Pi_1, \ldots, \Pi_W)$, is drawn. In this sample, each random $\Pi_w$ independently takes a value in $S$ according to the discrete uniform distribution over $S$. The Monte Carlo permutation p-value based on this sample is [Davison and Hinkley, 1997; Ernst, 2004; Phipson and Smyth, 2010]:

$$
\hat{p}^S_{t,W}(\Pi, G_{\text{obs}}, a) = (W + 1)^{-1} \left( \sum_{w=1}^{W} \left( t(G_{\text{obs}}, u_{\Pi_w}(a)) \geq t(G_{\text{obs}}, a) \right) + 1 \right)
$$

The Monte Carlo permutation test based on comparing the Monte Carlo permutation p-value to an arbitrary nominal $\alpha \in (0,1)$ rejects the null hypothesis with probability given by the test function:

$$
\phi^S_{t,\alpha,W}(\Pi, G_{\text{obs}}, a) = \begin{cases} 
1 & \hat{p}^S_{t,W}(\Pi, G_{\text{obs}}, a) \leq \alpha \\
\hat{p}^S_{t,W}(\Pi, G_{\text{obs}}, a) > \alpha 
\end{cases}
$$

(3-5)
If \( \mathbb{E}_F \left[ \hat{\phi}_{t,\alpha} (G_{\text{obs}}, a) | a, x \right] \leq \alpha \) for all \( t \) under a particular \( F \), then

\[
\mathbb{E}_F \left[ \hat{\phi}_{t,\alpha} (G_{\text{obs}}, a) | a, x \right] \leq \alpha \text{ for all } t \text{ under } F \text{ as well (see Subsection 3.7.2).}
\]

Similar results have been given previously with a different proof [Romano, 1989] or without proof [Lehmann and Romano, 2005]. Thus, conditions (I) and (II) together are sufficient to guarantee that the Monte Carlo permutation test based on \( S \) is exact level \( \alpha \). The subsequent sections evaluate whether these two conditions are both satisfied for specific Monte Carlo permutation tests in the presence of missing genotypes.

3.4 Standard Permutation Test

Let \( S_n \) refer to the set of the \( n! \) possible permutations of the integers. The standard permutation test, which is the most frequently used Monte Carlo permutation test for genetic case-control studies, is based on the test function in \((3-5)\), \( S = S_n \), and any choice of \( t \). \( S_n \) is a group under composition [Robinson, 2003], so condition (I) is trivially satisfied. The remainder of this section explores whether condition (II) is also satisfied for the standard permutation test.

The PMF \( f_{G,M} \) can be factored according to a selection model [Little and Rubin, 2002]:

\[
f_{G,M} ([\tilde{g}, \tilde{m}] | a, \tilde{x}, q) = f_G(\tilde{g} | a, \tilde{x}, q) f_M(\tilde{m} | \tilde{g}, a, \tilde{x}, q)
\]

The PMF \( f_G \) is the genotype sampling process, and \( f_M \) is the missing data process. The present study focuses on the impact of the missing data process on inference in the absence of other confounders such as population stratification. In such situations, the genotype sampling process does not depend on \( \tilde{x} \). Also assume that the distribution of
$\mathbf{M}$ does not depend intrinsically on disease status; that is, any dependency of $\mathbf{M}$ on $a$ arises only through the association, if any, of $a$ with $\mathbf{g}$ and $\mathbf{x}$. Thus, all possible data generating processes $\mathbb{F}$ are defined by (3-2) with an arbitrary $f_{G,M}$ having form:

$$f_{G,M}(\mathbf{g},\mathbf{m}|a,\mathbf{x},q) = f_G(\mathbf{g}|a,q)f_M(\mathbf{m}|\mathbf{g},\mathbf{x},q)$$ (3-6)

In a genetic case-control study, the genotype sampling process has the same form for all populations $q$:

$$f_G(\mathbf{g}|a,q) = \frac{f_A(a|\mathbf{g},q)f_G(\mathbf{g}|q)}{\sum_{g'}f_A(a|\mathbf{g},q)f_G(\mathbf{g}|q)}$$

where $f_A$ is the population disease model and $f_G$ is the unconditional genotype sampling process for population $q$. The null hypothesis specifies that the evolutionary process results in a population disease model that does not depend on the genotype at the locus of interest, which implies that $f_A(a|\mathbf{g},q) = f_A(a|q)$ and $f_G(\mathbf{g}|a,q) = f_G(\mathbf{g}|q)$ for all $q$. In other words, the genotype sampling process does not depend on affection status for all populations $q$ under the null hypothesis. Therefore, all possible data generating processes $\mathbb{F}$ consistent with the null hypothesis are defined by (3-2) with an arbitrary $f_{G,M}$ having form:

$$f_{G,M}(\mathbf{g},\mathbf{m}|a,\mathbf{x},q) = f_G(\mathbf{g}|q)f_M(\mathbf{m}|\mathbf{g},\mathbf{x},q)$$ (3-7)

Based on (3-4) and (3-7), condition (II) is satisfied if:

$$\prod_{k=1}^n \sum_{q^{-1}(\mathbf{g},\mathbf{x})} f_G(\mathbf{g}|q)f_M(\mathbf{m}|\mathbf{g},\mathbf{x},q) = \prod_{k=1}^n \sum_{q^{-1}(\mathbf{g},\mathbf{x})} f_G(\mathbf{g}|q)f_M(\mathbf{m}|\mathbf{g},\mathbf{x},q)$$ (3-8)
for all \( \pi \in S_n \), \( q \in Q \), and \( g_{\text{obs}} \in Z^n \). This condition is in turn satisfied if the PMF for individual \( \pi(k) \) is identical to the PMF for individual \( k \) for all \( \pi \in S_n \), \( q \in Q \), and \( k \), which requires only that \( \bar{x}_{\pi(k)} = \bar{x}_k \) for all \( \pi \in S_n \) and \( k \). However, if \( \bar{x}_{\pi(k)} \neq \bar{x}_k \) for some \( \pi \in S_n \) and \( k \), (3-8) may not hold for some \( g_{\text{obs}} \in Z^n \) and \( q \in Q \). Because there is no guarantee that \( \bar{x}_{\pi(k)} = \bar{x}_k \) for all \( \pi \in S_n \) and \( k \), condition (II) is not satisfied, and the standard permutation test is not guaranteed to be exact level \( \alpha \) in the presence of missing genotypes.

Suppose, however, that all possible \( \mathbb{F} \) arise from a missing data process that is independent of \( \bar{x} \). Under this condition, all possible \( \mathbb{F} \) consistent with the null hypothesis are defined by (3-2) with an arbitrary \( f_{G,M} \) having form:

\[
f_{G,M}(\bar{g}, \bar{m}|a, x, q) = f_G(g|q) f_M(m|g, q)
\]

Equation (3-4) is satisfied for such \( \mathbb{F} \) because:

\[
\prod_{k=1}^n \sum_{q^{-1}(\bar{g}_{\text{obs},k})} f_G(g|q) f_M(m|g, q) = \prod_{k=1}^n \sum_{q^{-1}(\bar{g}_{\text{obs},k})} f_G(g|q) f_M(m|g, q)
\]

for all \( \pi \in S_n \), \( q \in Q \), and \( g_{\text{obs}} \in Z^n \). Thus, the standard permutation test is level \( \alpha \) (i.e., has a type I error rate of at most \( \alpha \) for all \( t \) and all \( \mathbb{F} \) consistent with both the null hypothesis and the condition) when the missing data process is independent of \( \bar{x} \). Such missing data processes include complete data with \( f_M(m|g, q) = I[\bar{m} = 0, q] \) and randomly missing genotypes with \( f_M(m|g, q) = f_M(\bar{m}) \). Interestingly, this condition is also satisfied for missing genotypes that would be nonignorable for likelihood inference;
that is, the missing data process depends on the values of both observed and unobserved
genotypes but is conditionally independent of covariates [Little and Rubin, 2002].

The condition that the missing data process is independent of $\bar{x}$ is quite
restrictive. In practice, the rate and pattern of missing genotypes are likely to depend on
variables such as DNA source, age of sample, sequence capture method, sequencing
platform, reagent batch, genotype calling algorithm, and quality control parameters,
among others. While the standard permutation test is not necessarily invalid when the
missing data process depends on $\bar{x}$, it comes without a theoretically guaranteed level.
This lack of guaranteed level can be problematic; a Monte Carlo simulation subsequently
demonstrates that the standard permutation test may have highly inflated type I error rates
for some $t$ and $F$ consistent with the null hypothesis. Therefore, the following section
considers an alternative choice of $S$ to ensure that condition (II) is satisfied.

3.5 Restricted Permutation Test

Suppose that the elements of the covariate vector $\bar{x}$ are discrete or discretized
continuous variables and have been recorded completely for each individual. The various
values of these variables then define $C$ strata in the data set, and it can be assumed
without loss of generality that $\bar{x}$ is a scalar $x$ taking values in the integers $\{1,...,C\}$. Let
$I_c$ denote the set of indices $i$ for the $n_c$ individuals belonging to stratum $c$ and \( \hat{S}_n \subseteq S_n \)
be the set of all $n_1!\cdots n_c!$ permutations of the integers that exchange indices within but
not between $I_c$:

$$\hat{S}_n = \{ \pi \in S_n : \forall i \forall c, \pi(i) \in I_c \text{ if and only if } i \in I_c \}$$
The restricted permutation test is the Monte Carlo permutation test based on the test function in (3-5), \( S = \tilde{S}_n \), and any choice of \( t \).

The identity permutation, \( \pi_{id}(i) = i \), is an element of \( S_n \) that trivially satisfies the condition \( \pi_{id}(i) \in I_c \) if and only if \( i \in I_c \) for all \( i \) and \( c \), so \( \pi_{id} \in \tilde{S}_n \). For arbitrary \( \pi_1, \pi_2 \in \tilde{S}_n \), \( \pi_1, \pi_2 \in S_n \) because \( \tilde{S}_n \subseteq S_n \), so \( \pi_2 \circ \pi_1 \in S_n \). Also, \((\pi_2 \circ \pi_1)(i) = \pi_2(\pi_1(i)) \in I_c \) if and only if \( i \in I_c \) for all \( i \) and \( c \) because \( \pi_2(\pi_1(i)) \in I_c \) if and only if \( \pi_1(i) \in I_c \) and \( \pi_1(i) \in I_c \) if and only if \( i \in I_c \) for all \( i \) and \( c \). Thus, for all \( \pi_1, \pi_2 \in \tilde{S}_n \), \( \pi_2 \circ \pi_1 \in \tilde{S}_n \). Finally, for an arbitrary \( \pi \in \tilde{S}_n \), \( \pi^{-1} \in S_n \) because \( \tilde{S}_n \subseteq S_n \) and \( S_n \) is a group under composition [Robinson, 2003]. Because \( \pi(i) \in I_c \) if and only if \( i \in I_c \) for all \( i \) and \( c \), it follows that \( \pi(i) \in I_c \) if and only if \( \pi^{-1}(\pi(i)) \in I_c \) for all \( i \) and \( c \). Consequently, \( \pi^{-1}(i) \in I_c \) if and only if \( i \in I_c \) for all \( i \) and \( c \), which implies that \( \pi^{-1} \in \tilde{S}_n \) as well. Thus, for all \( \pi \in \tilde{S}_n \), \( \pi^{-1} \in \tilde{S}_n \). Based on the preceding, \( \tilde{S}_n \) is a subgroup of \( S_n \), which also means that it is a group under composition [Robinson, 2003]. Thus, condition (I) is satisfied for the restricted permutation test.

Consider once again all possible data generating processes \( \mathcal{F} \) consistent with the null hypothesis, which are defined by (3-2) with an arbitrary \( f_{G,M} \) having form given in (3-7). If \( x_k = c \) for an arbitrary \( k \), then \( k \in I_c \), \( \pi(k) \in I_c \) for an arbitrary \( \pi \in \tilde{S}_n \), and \( x_{\pi(k)} = c \) as well. For the converse, if \( x_{\pi(k)} = c \) for an arbitrary \( k \) and \( \pi \in \tilde{S}_n \), then \( \pi(k) \in I_c \), which implies that \( k \in I_c \) and \( x_k = c \). Therefore, the definition of \( \tilde{S}_n \) ensures...
that $x_{π(k)} = x_k$ for all $π \in \tilde{S}_n$ and $k$. From the previous section, it immediately follows that (3-8) is satisfied for all $π \in \tilde{S}_n$, $q \in Q$, and $g_{\text{obs}} \in Z^n$, which implies that condition (II) is satisfied and that the restricted permutation test is exact level $α$.

When one is unwilling to make assumptions regarding the missing data process, the restricted permutation test can always be used to perform permutation inference with a type I error rate of at most $α$. Because robustness generally requires sacrificing power, a logical question is whether the restricted permutation test is less powerful than the standard permutation test when the missing data process is independent of $\tilde{x}$ and both tests are level $α$. The answer is clear for one pathological case: each stratum contains only cases or controls. In this case, each observed genotype vector will have the same affection status indicator value in all restricted permutations, and the test statistic in each permutation will be the same as the test statistic in the observed data set. It follows that the p-value will be 1 for all $g_{\text{obs}} \in Z^n$, and the test will be exact level $α$ with zero power. While more general finite-sample theoretical results are difficult to obtain, a Monte Carlo simulation in the next section suggests that the reduction in power is probably modest and that other analytical choices may have substantially greater effects on power.

3.6 Monte Carlo Simulations

For each simulation, 1,000 replicates were generated from different $F$ defined by the same underlying evolutionary process and genotype sampling distribution but different missing data processes. To accomplish this, 1,000 replicate samples of genotypes $g$ were obtained through the two stage sampling procedure described in detail in Subsections 2.4.1-2.4.3. In summary, 1,000 replicate populations $q$ each comprising
10,000 haplotypes at a 100 kb locus were first generated based on a neutral coalescent model with recombination. For each \( q \), a balanced case-control sample \( g \) of total size 2,000 was then generated under random mating and a population-specific disease model determined according to general user-specified parameters. The genotype score \( G_j \) was the number of minor alleles in the genotype (0, 1, or 2), where the minor allele was determined based on the frequency in the haplotype population to ensure comparability across data sets from the same sample. The disease risk for an individual with no minor alleles at any risk variant was 5% in all simulations.

For each replicate \( g \), data sets \( g_{\text{obs}} \) were obtained from the same \( g \) under multiple missing data processes. This simulation design allowed us to examine the effect of the missing data process on inference conditional on the underlying sample of true genotypes. The missing data process \( f_M \) was modeled as follows. \( G_j \) and \( M_j \) for variant \( j \) in the random complete data vector for individual \( i \), \([\tilde{G}_i, \tilde{M}_i]\), are denoted by \( G_{ij} \) and \( M_{ij} \), respectively. The same notational convention applies to elements of \( \tilde{G}_{\text{obs},i} \).

Assuming that individuals were assayed in \( C \) batches over time, the conditional distribution of the missing genotype indicator for variant \( j \) in individual \( i \) assayed in batch \( c \) was modeled as:

\[
M_{ij} | g_{ij}, x_i = c, q \sim \text{Bernoulli}\left( \gamma_{0,c} + \gamma_{1,c} \left( g_{ij} - 2p_{jq} \right) \right)
\]

In this model, \( p_{jq} \) is the MAF for variant \( j \) in the population \( q \) from which \( g \) was drawn, and the \( \gamma_{c} \) are user-specified parameters for the missing data model in batch \( c \) applying to all variants, individuals, and populations. Each \( M_{ij} \) was obtained as an independent
draw from this conditional distribution, and the observed genotype $g_{obs,ij}$ was then set to $g_{ij}$ if $M_{ij} = 0$ or a missing code if $M_{ij} = 1$.

In a case-control sample from a randomly mating population $q$ under the null hypothesis, $f_{G_q}(g_{ij} | q) = \text{Binomial}(2, p_{jq})$ at each variant $j$ in each individual $i$ in each batch $c$. These draws are independent and identically distributed over individuals but not over variants within an individual due to linkage disequilibrium and variation in $p_{jq}$ across variants. Under this model for $M_{ij}$, random mating, and a null disease model, $\gamma_{0,c}$ can be interpreted as the expected rate of missing genotypes in batch $c$ for all data sets because:

$$
\begin{align*}
\mathbb{E}\left[M_{ij} \mid x_i = c, q\right] &= \mathbb{E}\left[\mathbb{E}\left[M_{ij} \mid G_{ij}, x_i = c, q\right] \mid q\right] = \mathbb{E}\left[\gamma_{0,c} + \gamma_{1,c} \left(G_{ij} - 2p_{jq}\right) \mid q\right] = \gamma_{0,c}
\end{align*}
$$

Regardless of the disease model, $\gamma_{1,c}$ is the linear change in the missing genotype probability with each additional minor allele in the true genotype for individuals in batch $c$.

The simulations considered $C=4$ with 450 cases/50 controls in Batch 0, 350 cases/150 controls in Batch 1, 150 cases/350 controls in Batch 2, and 50 cases/450 controls in Batch 3. This structure simulates a study in which the decision to genotype balanced numbers of cases and controls was made after large-scale genotyping of cases had already begun. Seven missing data processes summarized in Table 3.1 (Models 0-6) were considered. Model 0 corresponds to complete genotype data. For Models 1-6 with missing genotypes, the $\gamma_{0,c}$ were chosen so that, under a null disease model, the expected
rate of missing genotypes was $\sum_{c=0}^{3} \gamma_{0,c}/4 = 0.10$ in all data sets. Model 1 represents randomly missing genotypes. Model 2 results in nonrandomly missing genotypes and could arise if the same assay techniques were applied to all batches and the quality control measures preferentially set genotype calls containing minor alleles to missing. Models 3-6 also result in nonrandomly missing genotypes and could arise in studies in which each batch was genotyped with different assay techniques or at different laboratories.

Within each $g$, the same with-replacement samples of $W=10,000$ permutations from $S_n$ and $\tilde{S}_n$ were used for the Monte Carlo permutation tests with all $t$ and $g_{obs}$. However, these samples were independent for different $g$. All variants that were completely missing or monomorphic in each $g_{obs}$ were excluded from the analysis of that data set. For the restricted permutation test, the actual batch was used in defining $\tilde{S}_n$, which assumes that batch was identified as a potential factor influencing the missing data process and recorded on each subject.

Three different test statistics were considered. Chapter 2 established that inference based on efficiently combining single-variant Cochran-Armitage trend chi-square
statistics over variants in a locus is powerful in the presence of rare variants, so the CA max and CA sum tests were once again considered here. Let \( G_{\text{obs},j} = [G_{\text{obs},ij}, \ldots, G_{\text{obs},n}]^T \) denote the column of \( G_{\text{obs}} \) corresponding to variant \( j \) and:

\[
O_{ij} = I\left[G_{\text{obs},ij} \neq h\right]
\]

\[
N_{\text{obs},j} = \sum_{i=1}^{n} O_{ij}
\]

\[
\hat{\sigma}_{G_{\text{obs},j},a} = \left(N_{\text{obs},j} - 1\right)^{-1} \left[ \sum_{i=1}^{n} O_{ij} G_{\text{obs},ij} a_i - N_{\text{obs},j}^{-1} \sum_{i=1}^{n} O_{ij} G_{\text{obs},ij} \sum_{i=1}^{n} O_{ij} a_i \right]
\]

\[
\hat{\sigma}_{G_{\text{obs},j}}^2 = \left(N_{\text{obs},j} - 1\right)^{-1} \left[ \sum_{i=1}^{n} O_{ij} G_{\text{obs},ij}^2 - N_{\text{obs},j}^{-1} \left( \sum_{i=1}^{n} O_{ij} G_{\text{obs},ij} \right)^2 \right]
\]

\[
\hat{\sigma}_{a,j}^2 = \left(N_{\text{obs},j} - 1\right)^{-1} \left[ \sum_{i=1}^{n} O_{ij}^2 a_i^2 - N_{\text{obs},j}^{-1} \left( \sum_{i=1}^{n} O_{ij} a_i \right)^2 \right]
\]

The CA max statistic is defined as:

\[
t_{\text{CA max}}(G_{\text{obs}}, a) = \max_j \left\{ N_{\text{obs},j} \frac{\hat{\sigma}_{G_{\text{obs},j},a}^2}{\hat{\sigma}_{G_{\text{obs},j}}^2 \hat{\sigma}_{a,j}^2} \right\}
\]

The CA sum statistic is defined as:

\[
t_{\text{CA sum}}(G_{\text{obs}}, a) = \sum_{j=1}^{v} N_{\text{obs},j} \frac{\hat{\sigma}_{G_{\text{obs},j},a}^2}{\hat{\sigma}_{G_{\text{obs},j}}^2 \hat{\sigma}_{a,j}^2}
\]

Recall that the CA sum statistic is closely related to the SKAT [Wu et al., 2011] and C-alpha [Neale et al., 2011] statistics (see Subsections 2.4.5 and 2.5.3). The third statistic was a frequency-weighted burden (FW burden) statistic in the spirit of the WSS [Madsen and Browning, 2009]. For each data set, a Bayesian estimate of the MAF, \( \hat{p}_j \), among all individuals with observed genotypes and a frequency weight, \( \hat{w}_j \), were calculated for each variant \( j \) as follows:
\[ \hat{p}_j = \left( \sum_{i=1}^{n} O_{ij} G_{obs,ij} + 1 \right) / \left( 2N_{obs,ij} + 2 \right) \]

\[ \hat{w}_j = \left( 2 \hat{p}_j (1 - \hat{p}_j) \right)^{-1/2} \]

A genetic score \( \Gamma_i = \sum_{j=1}^{v} \hat{w}_j O_{ij} G_{obs,ij} \), which sums the frequency-weighted minor allele counts over all observed genotypes in each individual, was then calculated. Now let

\[ n_a = \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} a_i^2. \]

The FW burden statistic is the Wald chi-square statistic for the difference in mean genetic score between cases and controls, which is equivalently expressed as the product of \( n \) and the squared Pearson correlation between \( \Gamma_i \) and \( a_i \):

\[ t_{FW \text{ burden}} (G_{obs}, a) = \frac{\left[ n_a^{-1} \sum_{i=1}^{n} \Gamma_i a_i - (n - n_a)^{-1} \sum_{i=1}^{n} \Gamma_i (1 - a_i) \right]^2}{\left[ n_a^{-1} + (n - n_a)^{-1} \right] n^{-1} \left[ \sum_{i=1}^{n} \Gamma_i - n^{-1} \left( \sum_{i=1}^{n} \Gamma_i \right)^2 \right]} \]

The first simulation examined the effects of the missing data process on the type I error rate. Let \( F_{i,F,w}^S (p|\mathbf{a}, \mathbf{x}) = P_{\mathbb{F}} \left( \hat{P}_{i,w}^S (\mathbf{\Pi}, G_{obs}, \mathbf{a}) \leq p|\mathbf{a}, \mathbf{x} \right) \) denote the cumulative distribution function (CDF) of the Monte Carlo permutation p-value based on \( S \) and \( t \) under data generating process \( \mathbb{F} \). If \( \mathbb{F} \) is consistent with the null hypothesis, the type I error rate of the associated Monte Carlo permutation test at an arbitrary nominal \( \alpha \) is given by:

\[ \beta_{i,F,w}^S (\alpha) = E_{\mathbb{F}} \left[ \hat{P}_{i,F,w}^S (\mathbf{\Pi}, G_{obs}, \mathbf{a}) | \mathbf{a}, \mathbf{x} \right] = P_{\mathbb{F}} \left( \hat{P}_{i,w}^S (\mathbf{\Pi}, G_{obs}, \mathbf{a}) \leq \alpha | \mathbf{a}, \mathbf{x} \right) = F_{i,F,w}^S (\alpha | \mathbf{a}, \mathbf{x}) \]
If this Monte Carlo permutation test is level $\alpha$ under data generating process $\mathcal{F}$, then

$$ \beta_{t,F,W}^S(\alpha) \leq \alpha \text{ and } \beta_{t,F,W}^S(\alpha) - \alpha \leq 0 \text{ for all nominal } \alpha. $$

Based on the Glivenko-Cantelli Theorem [Wasserman, 2006], the empirical distribution function (EDF) of the above Monte Carlo permutation p-value over the 1,000 replicates from $\mathcal{F}$, $\hat{F}_{t,F,W}^S(\alpha|a,x)$, is a strongly consistent estimator of the function $\beta_{t,F,W}^S(\alpha)$. An exact 95% confidence envelope, $[\beta_{t,F,W}^S(\alpha), \beta_{t,F,W}^S(\alpha)]$, can also be obtained from the EDF using the Dvoretzky-Kiefer-Wolfowitz Inequality [Wasserman, 2006]. Therefore, $\hat{F}_{t,F,W}^S(\alpha|a,x) - \alpha$ is a strongly consistent estimator of and $[\beta_{t,F,W}^S(\alpha) - \alpha, \beta_{t,F,W}^S(\alpha) - \alpha]$ is an exact 95% confidence envelope for the function $\beta_{t,F,W}^S(\alpha) - \alpha$. This function is referred to as the type I error rate minus $\alpha$.

Figures 3.1-3.3 plot the estimated the type I error rate minus $\alpha$ and its 95% confidence envelope at 1,000 evenly spaced values of $-\log_{10}(\alpha)$ in $[0,4]$ for various $S$, $t$, and $\mathcal{F}$ consistent with the null hypothesis. The standard permutation test is level $\alpha$ when the missing data process is independent of $\bar{x}$. Only the missing data processes in Models 0-2 are independent of $\bar{x}$, and Figures 3.1-3.3 are consistent with the standard permutation test being level $\alpha$ under these conditions. However, condition (II) is not satisfied for the standard permutation test, and Figures 3.1-3.3 provide compelling evidence that it had a highly inflated type I error rate for some nominal $\alpha$ and test statistics with missing data processes in Models 3-6. Therefore, the standard permutation test is not exact level $\alpha$. The type I error inflation ranged from nonexistent to catastrophic depending on the particular combination of test statistic, missing data process, and nominal $\alpha$. For example, at a nominal $\alpha=0.01$, the range of estimated type I error rates for
missing data processes in Models 3-6 was [0.009, 0.022] for the CA max statistic, [0.009, 0.020] for the CA sum statistic, and a remarkably high [0.417, 0.955] for the FW burden statistic. Assuming that batch was correctly identified as a factor, recorded, and used in
defining \( \hat{S}_n \), condition (II) is satisfied for the restricted permutation test. Thus, the restricted permutation test is exact level \( \alpha \), and Figures 3.1-3.3 are consistent with this theoretical prediction.
The second simulation was performed to examine the difference in power between the standard and restricted permutation tests under various missing data processes for which both tests are level $\alpha$. The simulation used a logistic population disease model with 50 rare risk variants (representing ~5% of all variants in the locus in

**Figure 3.3: Type I error rates of Monte Carlo permutation tests based on the FW burden statistic**

The estimated (solid line) type I error rate minus $\alpha$ (vertical axis) and its 95% confidence envelope (gray ribbon) are plotted at 1,000 evenly spaced values of $-\log_{10}(\alpha)$ in [0,4] (horizontal axis). Estimates and confidence envelopes are based on 1,000 replicates from each $\mathbb{P}$ arising from a null disease model and each missing data process (Models 0-6); plots are faceted by test type (standard or restricted) and missing data process. The dashed black line and space below correspond to type I error rates less than or equal to $\alpha$. The second simulation was performed to examine the difference in power between the standard and restricted permutation tests under various missing data processes for which both tests are level $\alpha$. The simulation used a logistic population disease model with 50 rare risk variants (representing ~5% of all variants in the locus in
the average haplotype population) with population MAF<0.01 and OR=2. In each population \( q \), the risk variants were randomly chosen from among the variants meeting the MAF criterion. Table 3.2 shows the estimated power at a nominal \( \alpha \) of 0.05, 0.01, or 0.001 by test type and test statistic based on 1,000 replicates from various \( F \) arising from this disease model and different missing data processes in Models 0-2. The restricted permutation test did have lower power for all test statistics and nominal \( \alpha \), but the reduction was generally modest. This finding stands to reason because the standard and restricted permutation p-values for a given replicate, test statistic, and missing data process were correlated with one another (Figures 3.4-3.6). Importantly, power varied much more substantially with the choice of test statistic for the same

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>Missing Data Process</th>
<th>Test Type</th>
<th>CA max</th>
<th>CA sum</th>
<th>FW burden</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>Model 0</td>
<td>Standard</td>
<td>0.732</td>
<td>0.674</td>
<td>0.327</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Restricted</td>
<td>0.696</td>
<td>0.644</td>
<td>0.306</td>
</tr>
<tr>
<td></td>
<td>Model 1</td>
<td>Standard</td>
<td>0.727</td>
<td>0.677</td>
<td>0.323</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Restricted</td>
<td>0.675</td>
<td>0.633</td>
<td>0.297</td>
</tr>
<tr>
<td></td>
<td>Model 2</td>
<td>Standard</td>
<td>0.710</td>
<td>0.669</td>
<td>0.325</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Restricted</td>
<td>0.664</td>
<td>0.639</td>
<td>0.292</td>
</tr>
<tr>
<td>0.01</td>
<td>Model 0</td>
<td>Standard</td>
<td>0.574</td>
<td>0.448</td>
<td>0.204</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Restricted</td>
<td>0.533</td>
<td>0.436</td>
<td>0.169</td>
</tr>
<tr>
<td></td>
<td>Model 1</td>
<td>Standard</td>
<td>0.555</td>
<td>0.451</td>
<td>0.196</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Restricted</td>
<td>0.505</td>
<td>0.427</td>
<td>0.165</td>
</tr>
<tr>
<td></td>
<td>Model 2</td>
<td>Standard</td>
<td>0.559</td>
<td>0.447</td>
<td>0.200</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Restricted</td>
<td>0.508</td>
<td>0.431</td>
<td>0.161</td>
</tr>
<tr>
<td>0.001</td>
<td>Model 0</td>
<td>Standard</td>
<td>0.412</td>
<td>0.248</td>
<td>0.121</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Restricted</td>
<td>0.360</td>
<td>0.243</td>
<td>0.087</td>
</tr>
<tr>
<td></td>
<td>Model 1</td>
<td>Standard</td>
<td>0.391</td>
<td>0.247</td>
<td>0.120</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Restricted</td>
<td>0.347</td>
<td>0.244</td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td>Model 2</td>
<td>Standard</td>
<td>0.384</td>
<td>0.245</td>
<td>0.117</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Restricted</td>
<td>0.329</td>
<td>0.232</td>
<td>0.083</td>
</tr>
</tbody>
</table>
Figure 3.4: CA max restricted permutation p-values versus standard permutation p-values
CA max restricted permutation p-value (vertical axis) plotted against the corresponding standard permutation p-value (horizontal axis) for the same data set, both in \(-\log_{10}\) units. Points represent 1,000 replicates from various \(\mathcal{F}\) arising from a disease model with 50 rare risk variants (MAF<0.01; OR=2) and different missing data processes (Models 0-2); plots are faceted by missing data process.

Figure 3.5: CA sum restricted permutation p-values versus standard permutation p-values
CA sum restricted permutation p-value (vertical axis) plotted against the corresponding standard permutation p-value (horizontal axis) for the same data set, both in \(-\log_{10}\) units. Points represent 1,000 replicates from various \(\mathcal{F}\) arising from a disease model with 50 rare risk variants (MAF<0.01; OR=2) and different missing data processes (Models 0-2); plots are faceted by missing data process.
test type than it did with the choice of test type for the same test statistic under a given missing data process, which suggests that validity under all possible \( \mathcal{F} \) consistent with the null hypothesis rather than power should be the primary determinant of the choice of \( S \).

### 3.7 Detailed Derivations

#### 3.7.1 A Useful Property of a Permutation Test Function When \( S \) Is a Group under Composition

Suppose that \( S \) is a group under composition and consider an arbitrary \( t \) and \( g_{\text{obs}} \in \mathbb{Z}^n \). For an arbitrary \( \pi' \in S \), it follows from the definition of the permutation p-value that:

\[
p^S_t(u_{\pi'}(g_{\text{obs}}), a) = R^{-1} \sum_S I \left[ t\left( u_{\pi'}(u_{\pi'}(g_{\text{obs}})), a \right) \geq t(u_{\pi'}(g_{\text{obs}}), a) \right]
\]

\[
= R^{-1} \sum_S I \left[ t\left( u_{\pi'\pi}(g_{\text{obs}}), a \right) \geq t(u_{\pi'}(g_{\text{obs}}), a) \right]
\]

---

**Figure 3.6: FW burden restricted permutation p-values versus standard permutation p-values**

FW burden restricted permutation p-value (vertical axis) plotted against the corresponding standard permutation p-value (horizontal axis) for the same data set, both in \(-\log_{10}\) units. Points represent 1,000 replicates from various \( \mathcal{F} \) arising from a disease model with 50 rare risk variants (MAF<0.01; OR=2) and different missing data processes (Models 0-2); plots are faceted by missing data process.
Because \( S \) is a group under composition, \( \circ \) must be a binary operation on \( S \), which means that \( \pi' \circ \pi \in S \) for all \( \pi', \pi \in S \) [Robinson, 2003]. For an arbitrary \( \pi' \in S \), this binary operation defines a mapping \( \lambda_{\pi'} : S \to S \), where \( \lambda_{\pi'}(\pi) = \pi' \circ \pi \). Because \( S \) is a group under composition, \( (\pi')^{-1} \in S \) as well [Robinson, 2003], and \( \lambda_{(\pi')^{-1}} : S \to S \).

Composition of permutations in \( S \) must also satisfy the associative law [Robinson, 2003], so:

\[
\lambda_{\pi'} \left( \lambda_{(\pi')^{-1}}(\pi) \right) = \pi' \circ \left( (\pi')^{-1} \circ \pi \right) = \left( \pi' \circ (\pi')^{-1} \right) \circ \pi = \pi
\]

\[
\lambda_{(\pi')^{-1}} \left( \lambda_{\pi'}(\pi) \right) = (\pi')^{-1} \circ (\pi' \circ \pi) = \left( (\pi')^{-1} \circ \pi' \right) \circ \pi = \pi
\]

Therefore, \( \lambda_{(\pi')^{-1}} \) is the unique inverse for \( \lambda_{\pi'} \), which implies that \( \lambda_{\pi'} \) is a one-to-one mapping of \( S \) onto itself [Robinson, 2003]. Because \( \lambda_{\pi'} \) is one-to-one, different \( \pi \in S \) map to different \( \pi' \circ \pi \in S \), and the above sum contains \( R \) terms each corresponding to a distinct \( \pi' \circ \pi \in S \). Moreover, because any collection of \( R \) distinct elements of \( S \) must constitute the entire set, the above sum can equivalently be taken over all \( \pi \in S \). The permutation p-value for an arbitrary \( \pi' \in S \) can therefore be written as:

\[
p_{\pi'}^{(S)}(u_{\pi'}(g_{\text{obs}}), a) = R^{-1} \sum_{\pi} \mathbb{I} \left( t(u_{\pi'}(g_{\text{obs}}), a) \geq t(u_{\pi'}(g_{\text{obs}}), a) \right)
\]

Following previous developments [Hoeffding, 1952; Lehmann and Romano, 2005; Romano, 1989, 1990], consider the \( R \) ordered values of \( t(u_{\pi'}(g_{\text{obs}}), a) \) over all \( \pi \in S \):

\[
t^{(1)}(g_{\text{obs}}, a) \leq t^{(2)}(g_{\text{obs}}, a) \leq \cdots \leq t^{(R)}(g_{\text{obs}}, a)
\]
If $\alpha \in (0, R^{-1})$, it is clear that $p^S_i(u_\pi(g_{obs}), a) \geq R^{-1} > \alpha$ and $\phi^S_{i,\alpha}(u_\pi(g_{obs}), a) = 0$ for all $\pi' \in S$. This result implies that:

$$\sum_{\pi' \in S} \phi^S_{i,\alpha}(u_\pi(g_{obs}), a) = 0 < R\alpha$$

If $\alpha \in [R^{-1}, 1)$, $R\alpha$ can be separated into integral and decimal components according to $R\alpha = \lfloor R\alpha \rfloor + \varepsilon$, where $\lfloor \cdot \rfloor$ is the floor function and $\varepsilon \in [0,1)$ is a real number representing the decimal component. Define $r^* = R - \lfloor R\alpha \rfloor + 1$ and partition $S$ into three subsets:

$$S^+_i(g_{obs}, a) = \{\pi \in S : t(u_\pi(g_{obs}), a) > t(r^*) (g_{obs}, a)\}$$

$$S^0_i(g_{obs}, a) = \{\pi \in S : t(u_\pi(g_{obs}), a) = t(r^*) (g_{obs}, a)\}$$

$$S^-_i(g_{obs}, a) = \{\pi \in S : t(u_\pi(g_{obs}), a) < t(r^*) (g_{obs}, a)\}$$

For all $\pi' \in S^+_i(g_{obs}, a)$, there are at most $\lfloor R\alpha \rfloor - 1 \pi \in S$ for which $t(u_\pi(g_{obs}), a) \geq t(u_{\pi'}(g_{obs}), a)$, so $p^S_i(u_\pi(g_{obs}), a) \leq R^{-1}(\lfloor R\alpha \rfloor - 1) < \alpha$ and $\phi^S_{i,\alpha}(u_\pi(g_{obs}), a) = 1$. For all $\pi' \in S^0_i(g_{obs}, a)$, there are at least $\lfloor R\alpha \rfloor + 1 \pi \in S$ for which $t(u_\pi(g_{obs}), a) \geq t(u_{\pi'}(g_{obs}), a)$, so:

$$p^S_i(u_\pi(g_{obs}), a) \geq R^{-1}(\lfloor R\alpha \rfloor + 1) = R^{-1}(R\alpha - \varepsilon + 1) = \alpha + R^{-1}(1 - \varepsilon) > \alpha$$

and $\phi^S_{i,\alpha}(u_\pi(g_{obs}), a) = 0$. Finally, for all $\pi' \in S^0_i(g_{obs}, a)$, there are two possibilities. If $t(r^*) (g_{obs}, a) = t(r^{i-1}) (g_{obs}, a)$, then there are at least $\lfloor R\alpha \rfloor + 1 \pi \in S$ for which $t(u_\pi(g_{obs}), a) \geq t(u_{\pi'}(g_{obs}), a)$. Therefore, $p^S_i(u_\pi(g_{obs}), a) > \alpha$ and $\phi^S_{i,\alpha}(u_\pi(g_{obs}), a) = 0$ for all $\pi' \in S^0_i(g_{obs}, a)$, which implies that:
\[
\sum_{\pi' \in S_t} \phi^S_{t,\alpha}(g_{\text{obs}}, a) = \sum_{\pi' \in S_t'(g_{\text{obs}}, a)} 1 \leq \lfloor R\alpha \rfloor - 1 < R\alpha
\]

because there are at most \( \lfloor R\alpha \rfloor - 1 \) \( \pi' \in S_t'(g_{\text{obs}}, a) \) by definition. Alternatively, if 
\[
t^{(r')}(g_{\text{obs}}, a) > t^{(r'-1)}(g_{\text{obs}}, a),
\]
there must be exactly \( \lfloor R\alpha \rfloor \) \( \pi \in S \) with 
\[
t(u_{\pi}(g_{\text{obs}}, a) \geq t^{(r')}(g_{\text{obs}}, a).
\]
Therefore, 
\[
p^S_t(u_{\pi}(g_{\text{obs}}, a), a) = R^{-1}\lfloor R\alpha \rfloor \leq \alpha \quad \text{and}
\]
\[
\phi^S_{t,\alpha}(u_{\pi}(g_{\text{obs}}, a), a) = 1 \quad \text{for all} \quad \pi' \in S_t^0(g_{\text{obs}}, a).
\]
There are also exactly \( \lfloor R\alpha \rfloor \) 
\( \pi' \in S_t^0(g_{\text{obs}}, a) \cup S_t^+(g_{\text{obs}}, a) \), which implies that:
\[
\sum_{\pi' \in S_t} \phi^S_{t,\alpha}(u_{\pi}(g_{\text{obs}}, a), a) = \sum_{\pi' \in S_t'(g_{\text{obs}}, a)} 1 + \sum_{\pi' \in S_t^+(g_{\text{obs}}, a)} 1 = \lfloor R\alpha \rfloor \leq R\alpha
\]

Therefore, if \( S \) is a group under composition, then 
\( \sum_{S} \phi^S_{t,\alpha}(u_{\pi}(g_{\text{obs}}, a)) \leq R\alpha \) for all \( t \) and 
\( g_{\text{obs}} \in \mathbb{Z}^n \).

3.7.2 Relation between the Type I Error Rates of a Permutation Test and the Corresponding Monte Carlo Permutation Test

The finite-sample properties of the Monte Carlo permutation p-value and test have been examined previously for continuous data and test statistics using approximating uniform distributions for the true permutation p-value [Dwass, 1957; Phipson and Smyth, 2010]. However, the multi-variant genotypes and missing data indicators in genetic case-control studies are inherently discrete. Moreover, the true permutation p-value may not be well-approximated by a uniform random variable. While finite-sample results for Monte Carlo tests with discrete null distributions have been obtained previously [Jockel, 1986], these results apply to randomized tests rather than the nonrandomized test considered here. Finally, previous developments do not assume that
the Monte Carlo permutation p-value is calculated using the common computational shortcut of permuting the affection status indicators rather than the observed data vectors. Therefore, these previous developments are extended here to obtain a result suitable for the current application.

Suppose that a with-replacement random sample of $W \geq (1 - \alpha)/\alpha$ elements from $S$, $\Pi = (\Pi_1, \ldots, \Pi_W)$, is drawn. In this sample, each random $\Pi_w$ independently takes a value in $S$ according to the discrete uniform distribution over $S$. Note that $(W + 1)\alpha$ can be separated into integral and decimal components according to $(W + 1)\alpha =\lfloor (W + 1)\alpha \rfloor + \varepsilon$, where $\lfloor \cdot \rfloor$ is the floor function and $\varepsilon \in [0, 1)$ is a real number representing the decimal component. For an arbitrary matrix $g_{\text{obs}} \in \mathbb{Z}^n$, let $Y_w = I_1[t(g_{\text{obs}}, u_{\Pi_w}(a))] \geq t(g_{\text{obs}}, a)$ so that the Monte Carlo permutation p-value can be written as:

$$\hat{p}_{t,W}^S(\Pi, g_{\text{obs}}, a) = (W + 1)^{-1} \left( \sum_{w=1}^{W} Y_w + 1 \right)$$

For $\sum_{w=1}^{W} Y_w \leq \lfloor (W + 1)\alpha \rfloor - 1$:

$$\hat{p}_{t,W}^S(\Pi, g_{\text{obs}}, a) \leq (W + 1)^{-1} \left( \lfloor (W + 1)\alpha \rfloor \right) \leq \alpha$$

For $\sum_{w=1}^{W} Y_w \geq \lfloor (W + 1)\alpha \rfloor$:

$$\hat{p}_{t,W}^S(\Pi, g_{\text{obs}}, a) \geq (W + 1)^{-1} \left( \lfloor (W + 1)\alpha \rfloor + 1 \right)$$

$$= (W + 1)^{-1} \left( (W + 1)\alpha - \varepsilon + 1 \right)$$

$$= \alpha + (W + 1)^{-1}(1 - \varepsilon)$$

$$> \alpha$$
Therefore, \( \hat{\phi}^S_{\alpha W} (\Pi, g_{\text{obs}}, a) = 1 \) if and only if \( \sum_{w=1}^{W} Y_w \leq \left( (W + 1) \alpha \right) - 1 \). The rejection rate of the Monte Carlo randomization test for an arbitrary \( g_{\text{obs}} \in \mathbb{Z}^n \) can therefore be written as:

\[
E\left[ \hat{\phi}^S_{\alpha W} (\Pi, g_{\text{obs}}, a) \right] = P\left( \hat{\phi}^S_{\alpha W} (\Pi, g_{\text{obs}}, a) = 1 \right) = P\left( \sum_{w=1}^{W} Y_w \leq \left( (W + 1) \alpha \right) - 1 \right)
\]

The distribution of \( \sum_{w=1}^{W} Y_w \) must now be determined. Each random \( \Pi_w \) is independently drawn according to the discrete uniform distribution over \( S \), so:

\[
P(Y_w = 1) = P\left( t\left( g_{\text{obs}}, u_{\Pi_w} (a) \right) \geq t\left( g_{\text{obs}}, a \right) \right) = R^{-1} \sum_{S} I\left[ t\left( g_{\text{obs}}, u_{\pi} (a) \right) \geq t\left( g_{\text{obs}}, a \right) \right]
\]

Because \( t\left( G_{\text{obs}}, a \right) = t\left( [\tilde{G}_{\text{obs},1}, a_1], \ldots, [\tilde{G}_{\text{obs},n}, a_n] \right) \) is symmetric in its vector arguments \( [\tilde{G}_{\text{obs},i}, a_i] \), it follows for an arbitrary \( \pi \in S \) that:

\[
t\left( g_{\text{obs}}, u_{\pi} (a) \right) = t\left( [\tilde{g}_{\text{obs},1}, a_{\pi(1)}], \ldots, [\tilde{g}_{\text{obs},n}, a_{\pi(n)}] \right) = t\left( [\tilde{g}_{\text{obs}, \pi^{-1}(1)}, a_1], \ldots, [\tilde{g}_{\text{obs}, \pi^{-1}(n)}, a_n] \right) = t\left( u_{\pi^{-1}} (g_{\text{obs}}), a \right)
\]

where the second equality results from sorting the vector arguments so that the affection status indicators appear in the same order as in the original data set. The above probability can therefore be equivalently expressed as:

\[
P(Y_w = 1) = R^{-1} \sum_{S} I\left[ t\left( u_{\pi^{-1}} (g_{\text{obs}}), a \right) \geq t\left( g_{\text{obs}}, a \right) \right]
\]
Consider the mapping $\lambda(\pi) = \pi^{-1}$ that takes an arbitrary permutation $\pi$ to its unique inverse permutation $\pi^{-1}$. Because $S$ is a group under composition, $\pi^{-1} \in S$ for all $\pi \in S$ [Robinson, 2003], so $\lambda : S \to S$. Because $\pi$ is also by definition the unique inverse for $\pi^{-1}$, it follows that $\lambda(\lambda(\pi)) = \lambda(\pi^{-1}) = \pi$ for all $\pi \in S$. Therefore, $\lambda$ is its own unique inverse, which implies that $\lambda$ is a one-to-one mapping of $S$ onto itself [Robinson, 2003].

Because $\lambda$ is one-to-one, different $\pi \in S$ map to different $\pi^{-1} \in S$, and the above sum contains $R$ terms each corresponding to a distinct $\pi^{-1} \in S$. Moreover, because any collection of $R$ distinct elements of $S$ must constitute the entire set, the above sum can equivalently be taken over all $\pi \in S$. This result implies that:

$$P(Y_w = 1) = R^{-1} \sum_{S} I \left[ t\left(\mu_\pi (g_{\text{obs}}), a\right) \geq t(g_{\text{obs}}, a) \right]$$

$$= p_i^S (g_{\text{obs}}, a)$$

Thus, for an arbitrary $g_{\text{obs}} \in \mathbb{Z}^n$, each $Y_w$ is an independent Bernoulli($p_i^S (g_{\text{obs}}, a)$) random variable, and it immediately follows that the distribution of $\sum_{w=1}^{W} Y_w$ is Binomial($W, p_i^S (g_{\text{obs}}, a)$). Based on the preceding results:

$$E\left[ \hat{\phi}_{i,\alpha, W} (\Pi, g_{\text{obs}}, a) \right] = \sum_{y=0}^{\lfloor (W+1)\alpha \rfloor - 1} \binom{W}{y} (p_i^S (g_{\text{obs}}, a))^y (1 - p_i^S (g_{\text{obs}}, a))^{W-y}$$

$$= \psi_{\alpha, W} (p_i^S (g_{\text{obs}}, a))$$

The law of iterated expectations yields an expression for the unconditional rejection rate of the Monte Carlo randomization test under a data generating process $\mathbb{F}$:

$$E_{\mathbb{F}} \left[ \hat{\phi}_{i,\alpha, W} (\Pi, G_{\text{obs}}, a) \bigg| a, x \right] = E_{\mathbb{F}} \left[ E_{\mathbb{F}} \left[ \hat{\phi}_{i,\alpha, W} (\Pi, g_{\text{obs}}, a) \bigg| G_{\text{obs}} = g_{\text{obs}} \right] a, x \right]$$

$$= E_{\mathbb{F}} \left[ \psi_{\alpha, W} (p_i^S (G_{\text{obs}}, a)) \bigg| a, x \right]$$
The well-known relationship between the binomial survival function and the incomplete beta function [Zelen and Severo, 1965] can be used to obtain an alternative expression for \( \psi_{\alpha,W}(p) \) similar to one given by Jockel [Jockel, 1986]:

\[
\psi_{\alpha,W}(p) = 1 - \sum_{y=\lfloor (W+1)\alpha \rfloor}^{W} \binom{W}{y} p^y (1 - p)^{W-y}
= 1 - \beta\left(\left\lfloor (W+1)\alpha \right\rfloor, W - \left\lfloor (W+1)\alpha \right\rfloor + 1\right) - \int_0^p z^{\left\lfloor (W+1)\alpha \right\rfloor - 1} (1 - z)^{W - \left\lfloor (W+1)\alpha \right\rfloor} \, dz
\]

where \( \beta(a,b) = \int_0^1 z^{a-1} (1 - z)^{b-1} \, dz \) is the beta function [Davis, 1965]. Note that \( \psi_{\alpha,W}(1) = 0 \) by the definition of the beta function. Based on the relationships between the beta function, the gamma function, and factorials [Davis, 1965], the above expression can be further simplified:

\[
\psi_{\alpha,W}(p) = 1 - \frac{\Gamma(W+1)}{\Gamma(W - \left\lfloor (W+1)\alpha \right\rfloor + 1)} \frac{\Gamma\left(\left\lfloor (W+1)\alpha \right\rfloor \right)}{\Gamma\left(\left\lfloor (W+1)\alpha \right\rfloor + 1\right)} \int_0^p z^{\left\lfloor (W+1)\alpha \right\rfloor - 1} (1 - z)^{W - \left\lfloor (W+1)\alpha \right\rfloor} \, dz
= 1 - \frac{W!}{(W - \left\lfloor (W+1)\alpha \right\rfloor)!}\frac{\Gamma\left(\left\lfloor (W+1)\alpha \right\rfloor \right)}{\Gamma\left(\left\lfloor (W+1)\alpha \right\rfloor + 1\right)} \int_0^p z^{\left\lfloor (W+1)\alpha \right\rfloor - 1} (1 - z)^{W - \left\lfloor (W+1)\alpha \right\rfloor} \, dz
= 1 - W\left(\left\lfloor (W+1)\alpha \right\rfloor - 1\right) \int_0^p z^{\left\lfloor (W+1)\alpha \right\rfloor - 1} (1 - z)^{W - \left\lfloor (W+1)\alpha \right\rfloor} \, dz
\]

This expression for \( \psi_{\alpha,W}(p) \) has been obtained previously by Dwass [Dwass, 1957].

Let \( P \sim \text{Uniform}(0,1) \) and suppose that \( E_{\mathcal{F}}[\phi_{t,\alpha}^S(G_{\text{obs}}, a) | a, x] \leq \alpha \) for all \( t \) under a particular \( \mathcal{F} \). This assumption implies that:

\[
P_{\mathcal{F}}\left(p_i^S(G_{\text{obs}}, a) \leq \alpha | a, x\right) = E_{\mathcal{F}}[\phi_{t,\alpha}^S(G_{\text{obs}}, a) | a, x] \leq \alpha
\]

for all \( t \) under \( \mathcal{F} \). Because the inequality holds for an arbitrary nominal \( \alpha \), it holds for all nominal \( \alpha \in (0,1) \), and it immediately follows that \( P_{\mathcal{F}}\left(p_i^S(G_{\text{obs}}, a) \leq p | a, x\right) \leq p \) and
$$P_F\left(p_t^s(G_{obs}, a) > p | a, x\right) \geq 1 - p = P(P > p) \text{ for all } p \in (0, 1) \text{ and } t \text{ under } \mathbb{F}.$$ Therefore, the random variable $p_t^s(G_{obs}, a)$ is stochastically larger than $P$ [Ross, 1996] for all $t$ under $\mathbb{F}$. Moreover, the function $1 - \psi_{\alpha, W}(p)$ is increasing in $p$ for all $p \in [0, 1]$ because:

$$\frac{d}{dp} \left(1 - \psi_{\alpha, W}(p)\right) = W \left(\frac{W - 1}{\lfloor (W + 1)\alpha \rfloor - 1}\right) p^{\lfloor (W + 1)\alpha \rfloor - 1} (1 - p)^{W - \lfloor (W + 1)\alpha \rfloor} \geq 0$$

It follows that, for all $t$ under $\mathbb{F}$ [Ross, 1996]:

$$E_F\left[1 - \psi_{\alpha, W}(p_t^s(G_{obs}, a)) | a, x\right] \geq E\left[1 - \psi_{\alpha, W}(P)\right]$$

$$1 - E_F\left[\psi_{\alpha, W}(p_t^s(G_{obs}, a)) | a, x\right] \geq 1 - E\left[\psi_{\alpha, W}(P)\right]$$

$$E_F\left[\psi_{\alpha, W}(p_t^s(G_{obs}, a)) | a, x\right] \leq E\left[\psi_{\alpha, W}(P)\right]$$

The expectation $E\left[\psi_{\alpha, W}(P)\right]$ can be calculated using integration by parts [Dwass, 1957; Jockel, 1986]:

$$E\left[\psi_{\alpha, W}(P)\right] = \int_0^1 \psi_{\alpha, W}(p) \, dp$$

$$= \left[\psi_{\alpha, W}(p) \right]_0^1 - \int_0^1 p \psi'_{\alpha, W}(p) \, dp$$

$$= \psi_{\alpha, W}(1) - \psi_{\alpha, W}(0) + W \left(\frac{W - 1}{\lfloor (W + 1)\alpha \rfloor - 1}\right) \int_0^1 p^{\lfloor (W + 1)\alpha \rfloor - 1} (1 - p)^{W - \lfloor (W + 1)\alpha \rfloor} \, dp$$

$$= W \left(\frac{W - 1}{\lfloor (W + 1)\alpha \rfloor - 1}\right) B\left(\lfloor (W + 1)\alpha \rfloor + 1, W - \lfloor (W + 1)\alpha \rfloor + 1\right)$$

$$= \frac{W! \Gamma\left(\lfloor (W + 1)\alpha \rfloor + 1\right) \Gamma\left(W - \lfloor (W + 1)\alpha \rfloor + 1\right)}{\lfloor (W + 1)\alpha \rfloor - 1)! (W - \lfloor (W + 1)\alpha \rfloor) ! \Gamma(W + 2)}$$

$$= \frac{W! \Gamma\left(\lfloor (W + 1)\alpha \rfloor + 1\right) (W - \lfloor (W + 1)\alpha \rfloor) ! (W + 1)!}{\lfloor (W + 1)\alpha \rfloor - 1)! (W - \lfloor (W + 1)\alpha \rfloor) ! (W + 1)!}$$

$$= \frac{\lfloor (W + 1)\alpha \rfloor}{(W + 1)}$$

$$\leq \alpha$$
Based on the preceding results, if \( E_\mathcal{Y} \left[ \phi_{t,\alpha}^S (G_{\text{obs}}, a) \mid a, x \right] \leq \alpha \) for all \( t \) under a particular \( \mathbb{P} \),

then:

\[
E_\mathcal{Y} \left[ \hat{\phi}_{t,\alpha^\ast}^S (\Pi, G_{\text{obs}}, a) \mid a, x \right] = E_\mathcal{Y} \left[ \psi_{\alpha^\ast}^S (P_{t}^S (G_{\text{obs}}, a)) \mid a, x \right] \\
\leq E \left[ \psi_{\alpha^\ast}^S (P) \right] \\
\leq \alpha
\]

for all \( t \) under \( \mathbb{P} \) as well.
CHAPTER 4: DISCUSSION

4.1 Summary of Findings

Three existing methods for locus-wide inference using nonnegative single-variant test statistics were compared to two widely cited pooling tests in terms of their ability to detect associations between rare variants and disease. A simple model of a locus with one rare risk and one rare neutral variant, combined with data on the characteristics of variants in actual sequence data, demonstrated that even using Bonferroni-corrected single-variant tests for locus-wide inference may have higher power than collapsing or summing rare variant minor alleles in the presence of a neutral variant. To extend this conclusion, populations of realistic haplotypes at a hypothetical 100 kb locus were simulated, and power was examined for balanced case-control samples drawn from these populations according to a disease model with heterogeneous risk alleles and extensive neutral variation. In these simulations, one or more of the existing approaches for efficient locus-wide inference using nonnegative single-variant test statistics, the CA max test or CA sum test, had power comparable to or greater than the CMC and WSS tests under the scenarios considered. Moreover, the type I error and power of the CA max and CA sum tests were robust to randomly missing genotype data, which was not observed with the CMC test. Finally, the CA max test was nearly always more powerful than the CA sum test for disease models with only rare risk variants, suggesting that the CA max

---

3 Portions of this chapter have been published previously as: Kinnamon DD, Hershberger RE, Martin ER. 2012. Reconsidering association testing methods using single-variant test statistics as alternatives to pooling tests for sequence data with rare variants. PLoS ONE 7:e30238.
test may outperform the class of techniques represented by the SKAT and C-alpha test in these scenarios.

Because methods for efficient locus-wide inference using nonnegative single-variant test statistics may rely on permutation tests, a method for constructing exact level- \( \alpha \) permutation tests was also developed for genetic case-control studies with missing genotypes. To construct a Monte Carlo permutation test, one can simply choose a set of permutations and a test statistic. Provided that two conditions are satisfied by these choices, the resulting Monte Carlo permutation test is exact level \( \alpha \) in the presence of missing genotypes. One of these conditions is not satisfied for the standard permutation test, and a Monte Carlo simulation demonstrated that it is not exact level \( \alpha \) and may have extremely inflated type I error for some test statistics and missing data processes. However, if the missing data process is influenced by known covariates that are recorded, subjects can be divided into strata, and the set of permutations that shuffles affection status indicators only within strata can be used. The resulting test, which is termed the restricted permutation test, is exact level \( \alpha \) in the presence of missing genotypes. This theoretical prediction was confirmed by a Monte Carlo simulation, which demonstrated that the restricted permutation test can recover control of type I error lost by the standard permutation test. A second Monte Carlo simulation suggested that the reduction in power from using the restricted permutation test is generally modest when both tests are level \( \alpha \). In fact, the choice of test statistic had a much greater impact on power than the choice of test type.
4.2 Addressing the Challenge Posed by Rare Variants

The results of the present study contradict those of the original studies [Li and Leal, 2008; Madsen and Browning, 2009] suggesting that the CMC and WSS tests were superior to locus-wide inference using nonnegative single-variant test statistics. However, the simulations presented in Section 2.4 improve upon these studies in two important ways that explain the differences in results and make the results presented here more relevant to the analysis of actual sequence data. First, a widely accepted population genetic model, the coalescent, was used to simulate variants with MAF and LD distributions similar to those in actual sequence data, meaning that these simulations should more accurately reflect the impact of neutral variants on each method. Although the CMC study did consider the impact of including neutral variants, it used analytic power calculations that assumed independence between genotypes at different variant sites [Li and Leal, 2008]. The study also considered only models with fixed numbers of variants of different types having equal MAFs within each type. The WSS study considered only MAF spectra consistent with mildly deleterious mutations and sampled each variant, whether risk or neutral, independently of all others [Madsen and Browning, 2009]. Neither of these methods is likely to recapitulate the rich complexity of the variant MAF and LD distributions observed in actual sequence data as well as the coalescent-based approach in the present study did. Moreover, simulated data without many higher-frequency neutral variants or substantial LD between neutral and risk variants would tend to cause fewer problems with noise and masking in pooling tests, resulting in overly optimistic assessments of the performance of these techniques. In fact, the WSS test was often less powerful than even the inefficient BC-CA test, suggesting that noise and
masking from neutral variants may present major problems for techniques based on summing in actual sequence data.

Second, the present study compared pooling tests to efficient methods for locus-wide inference based on nonnegative single-variant test statistics that reduce the multiple-testing penalty by accounting for LD-induced correlations between the single-variant test statistics. However, the CMC and WSS tests were both compared in the original studies to the Bonferroni and Dunn-Sidak corrections [Li and Leal, 2008; Madsen and Browning, 2009], which are both generally conservative. Although the choice to assume independence between variants in the original WSS study should mean that the Dunn-Sidak correction was efficient, the Bonferroni correction used in the original CMC study should still have been conservative and thus inefficient under these conditions. In the more realistic simulated data used in the present study, LD would have induced correlations between test statistics at different variants, which would have rendered both of these techniques more conservative [Westfall and Young, 1993]. In such situations, methods based on simulating the joint distribution of p-values or test statistics under the locus-wide null hypothesis yield more powerful locus-wide tests [Westfall and Young, 1993] and are the relevant targets for comparison. The CA max test is one such method, and it outperformed the BC-CA test under every scenario considered in the simulations in Section 2.4 while controlling the Type I error rate, as predicted by theory. Thus, the CA max test, which is simple and computationally feasible, provides a fairer representation of the performance of existing methods for efficient locus-wide inference using nonnegative single-variant test statistics in actual sequence data.
The problem of neutral and protective rare variants masking case-control differences in pooling tests has been recognized by other authors [Hoffmann et al., 2010; Ionita-Laza et al., 2011; Neale et al., 2011; Price et al., 2010; Sul et al., 2011; Wu et al., 2011]. Many new developments have therefore sought to reduce the influence of putative neutral and protective variants using filtering, classification, or weights based on annotation, functional predictions, or MAFs [Hoffmann et al., 2010; Ionita-Laza et al., 2011; Price et al., 2010; Sul et al., 2011; Wu et al., 2011]. While these approaches seem sensible, there are several drawbacks. First, annotation and functional predictions are not readily available for non-coding sequences that may influence disease through recently discovered or as-yet-unknown regulatory mechanisms. Second, as demonstrated by recent examples implicating synonymous coding variants in altered protein products and Crohn’s disease [Brest et al., 2011; Kimchi-Sarfaty et al., 2007], annotation and functional predictions for coding sequences do not always provide a solid basis on which to separate putative risk, neutral, and protective variants a priori. Finally, distinguishing neutral and protective variants based on sample MAFs alone [Ionita-Laza et al., 2011] will be prone to error because of sampling variability, particularly with rare variants. In contrast, methods for locus-wide inference using nonnegative single-variant test statistics are inherently robust to the inclusion of neutral and protective variants and may even be able to exploit their LD with risk variants to increase power. Notably, the power advantage of the CA max and CA sum tests observed in the present study did not require any information or assumptions about the putative functional consequences of the minor allele in relation to the disease of interest. Thus, the CA max or CA sum tests could be
applied equally well to coding sequence, non-coding sequence with poorly understood functional consequences, or a combination thereof.

Methods for locus-wide inference using nonnegative single-variant test statistics can be used to address the allelic heterogeneity due to rare variants in more general association testing problems. Although the present study focused on single-locus inference for concreteness, test statistics can be combined over any relevant grouping of variants, including single exons, pathways, or the entire exome, to perform joint inference. Pooling tests can also be applied to arbitrary groupings, but they are not inherently robust to the inclusion of neutral and protective variants. Moreover, although the present study focused on case-control association testing in the absence of confounding and population stratification, existing methods using nonnegative single-variant test statistics can be readily extended to multi-variant joint inference in more complex case-control or family-based designs by simply changing the test statistic and permutation strategy. As long as the new test statistic has a nonnegative value that depends only on the magnitude of the deviation from the null hypothesis at each variant, the locus-wide test is inherently robust to the inclusion of neutral and protective variants. The permutation strategy would then need to be adapted to ensure exchangeability under the model implied by the new single-variant test statistic (see, e.g., [McIntyre et al., 2000] for a permutation strategy valid for the transmission/disequilibrium test statistic in a trio design). Finally, although only the maximum and sum of Cochran-Armitage trend chi-square statistics over the variant grouping of interest were considered in the present study, almost any summary of a wide variety of nonnegative single-variant test statistics could be used for joint inference based on the appropriate permutation distribution.
Although the idea of pooling minor alleles in association tests with rare variants may still hold sway in the genetics community, it is worth noting that some new association tests with greater robustness to the inclusion of neutral and protective variants have implicitly returned to locus-wide inference using nonnegative single-variant test statistics. Specifically, the SKAT [Wu et al., 2011] and C-alpha test [Neale et al., 2011] are equivalent to basing inference on weighted and unweighted sums of squared single-variant score statistics, respectively [Wu et al., 2011]. The sum $T$ statistic is also a sum of squared single-variant score statistics each weighted by the inverse of its estimated null variance. The results for the CA sum test in the present study, combined with the results of the studies proposing the SKAT and C-alpha test [Neale et al., 2011; Wu et al., 2011], suggest that further extending methods for locus-wide inference using nonnegative single-variant test statistics may be a fruitful line of research. Moreover, a method in this class fundamentally different from the closely related SKAT, C-alpha test, and CA sum test—the CA max test—often had greater power than the CA sum test for disease models with only rare risk variants. The results of the present study therefore suggest that a conceptual framework based on optimally combining nonnegative single-variant test statistics may yield useful insights or suggest other existing techniques that might be overlooked within a conceptual framework based on pooling minor alleles.

### 4.3 Addressing the Challenge Posed by Missing Genotypes

The present study showed that existing methods that use permutation inference to efficiently combine nonnegative single-variant test statistics are robust to randomly missing genotypes. This robustness stands in stark contrast to the observations for the CMC test using Hotelling’s $T^2$, which rapidly became unreliable with as little as 0.5%
randomly missing genotypes. Other multivariate techniques that rely on a generalized linear model framework, such as the SKAT, will also be subject to the same problem because generalized linear models can only use individuals with complete data. Although all individuals’ data could be made complete by imputing missing genotypes, low-frequency or rare variants may be difficult to impute with high accuracy (see, e.g., [Li et al., 2011]).

Although methods that use permutation inference to efficiently combine nonnegative single-variant test statistics are robust to randomly missing genotypes, caution must be exercised when constructing permutation tests for more general missing data processes. The frequently used standard permutation test is not exact level $\alpha$, and using it requires assuming that the missing data process is independent of observed and unobserved covariates. If this assumption cannot be justified and all covariates influencing the missing data process have been identified and completely recorded, an exact level $\alpha$ permutation test can be constructed by permuting only within strata defined by the levels of these discrete or discretized covariates. Notably, these results apply to any association test statistic and, therefore, to any association test for genetic case-control studies that relies on permutation inference.

While this modification of the permutation procedure can prevent inflated type I error, the design of data collection and assaying procedures at the beginning of the study can still play an important role in determining power. The standard permutation test is not guaranteed to be exact level $\alpha$ for designs in which the values of covariates influencing the missing data process are perfectly separated by affection status. One such design is the shared controls design, in which all control genotypes are obtained from an existing
data set or a publicly available database. Unfortunately, the restricted permutation test also has zero power for these designs, and it can easily be seen that there is no set of permutations that can provide a test with non-zero power for which conditions (I) and (II) in Section 3.3 are both satisfied. Therefore, the usual benefits of permutation inference may not apply in shared controls designs, and this fact should be carefully considered before undertaking such a design.

In contrast, designs in which both controls and cases are assayed concurrently in batches allow for valid permutation inference even if the missing data process changes substantially between batches due to changing DNA sources, genotyping platforms, reagents, and quality control measures. The restricted permutation test is exact level $\alpha$ with missing genotypes for these types of designs, and Monte Carlo simulation results suggested that its power is not necessarily substantially lower than the standard permutation test even when the latter is valid.

Other studies have shown that the asymptotic validity of permutation tests depends heavily on the choice of test statistic when the group invariance assumption is not met due to factors other than missing data [Chung and Romano, 2011; Romano, 1990]. Such results offer clues as to why the type I error inflation for the standard permutation test varied substantially across test statistics in the Monte Carlo simulation in Section 3.6. Because the choice to use permutation tests is often made when asymptotic approximations are questionable, the best initial course of action may be to modify the permutation procedure so that the group invariance assumption is satisfied in the presence of missing genotypes. However, such modifications are not always possible due to factors such as study design, so identifying test statistics that yield asymptotically valid and
powerful permutation tests in the presence of missing genotypes could be a productive line of future inquiry.

In a broader sense, the results in Chapter 3 also suggest that permutation tests can offer simple nonparametric alternatives to likelihood-based parametric procedures for analyzing genetic case-control studies in which the missing genotypes are nonignorable for likelihood inference [Little and Rubin, 2002]. These situations arise when the missing data process depends on the true values of the unobserved genotypes, and valid likelihood inference requires correct modeling of both the genotype sampling process and the missing data process as well as assumptions on their parameter spaces [Little and Rubin, 2002]. Even with correct model specification, the high dimensionality of genetic data makes model identification difficult. In contrast, the standard permutation test is level $\alpha$ with nonignorable missing genotypes if the missing data process depends on the true genotypes in the same way in all individuals. Moreover, the restricted permutation test is exact level $\alpha$ and can be used as long as the values of all covariates influencing the missing data process were recorded for all individuals. Permutation tests may therefore facilitate valid inference in otherwise intractable situations with nonignorable missing genotypes.

In closing, it is important to note that the two sufficient conditions provided in Section 3.3 can be applied more broadly to constructing exact level-$\alpha$ permutation tests that address problems such as population stratification in genetic case-control studies with missing genotypes. While others have proposed a permutation-like procedure to address problems such as population stratification [Epstein et al., 2012], their approach requires maximum likelihood estimation of a parametric model of the odds of disease
conditional on potential confounding variables. For this reason, their procedure results in parametric bootstrap tests, which have type I error rates only asymptotically approaching the nominal $\alpha$ [Efron and Tibshirani, 1993; Romano, 1989]. In contrast, any Monte Carlo permutation test satisfying the two sufficient conditions provided in Section 3.3 is fully nonparametric and exact level $\alpha$. To construct a Monte Carlo permutation test, one would begin by defining a testing problem based on a set of assumptions regarding the genotype sampling process and missing data process in the nonparametric probability model from Section 3.2. One could then define a set of permutations and obtain an exact level-$\alpha$ permutation test by verifying that conditions (I) and (II) in Section 3.3 are both satisfied. Given that the set of permutations yields an exact level-$\alpha$ permutation test, one could finally select a test statistic with high power under the suspected alternative based on considerations discussed elsewhere [Hoeffding, 1952; Lehmann and Romano, 2005; Lehmann and Stein, 1949; Pesarin and Salmaso, 2010].
REFERENCES


